LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

UCRL-TR-222758

Subspace Detectors: Theory

D. B. Harris

July 11, 2006

This work was performed under the auspices of the U.S. Department of Energy by University of California, Lawrence Livermore National Laboratory under Contract W-7405-Eng-48.

Subspace Detectors: Theory

Dave Harris
Lawrence Livermore National Laboratory
July 14, 2006

Abstract

Broadband subspace detectors are introduced for seismological applications that require the detection of repetitive sources that produce similar, yet significantly variable seismic signals. Like correlation detectors, of which they are a generalization, subspace detectors often permit remarkably sensitive detection of small events. The subspace detector derives its name from the fact that it projects a sliding window of data drawn from a continuous stream onto a vector signal subspace spanning the collection of signals expected to be generated by a particular source. Empirical procedures are presented for designing subspaces from clusters of events characterizing a source. Furthermore, a solution is presented for the problem of selecting the dimension of the subspace to maximize the probability of detecting repetitive events at a fixed false alarm rate. An example illustrates subspace design and detection using events in the 2002 San Ramon, California earthquake swarm.

Introduction

Motivation

Correlation detectors are gaining in popularity due to the significant reductions in detection threshold that are sometimes possible [Wiechecki-Vergara et al., 2001; Gibbons and Ringdal, 2006] and due to the ability of such detectors to classify events as they are detected [Harris, 1991, 1997, 2001]. Correlation detectors applied to array signals solve an age-old problem of beam-forming loss due to signal decorrelation across a receiver aperture [Mykkeltveit et al., 1983], but at the expense of introducing a new problem of signal decorrelation in the source region. Subspace detectors [Scharf and Friedlander, 1994; Harris 1997, 2001, 2003, 2004] offer a partial solution to this problem by extending the space of signals that can be detected from a space spanned by a single waveform template to a subspace chosen to span the range of signals observed from previous events occurring at a source of interest.

The detection and identification of proximate events producing similar signals occurs in many seismological applications. For example, repeating events are used as markers along faults to estimate rates of slip and energy release [Nadeau et al., 1995; Nadeau and McEvilly, 1997] and to estimate local changes in seismic velocity induced by large earthquakes [Schaff and Beroza, 2004]. The proximity of earthquakes and mining explosions has been estimated using waveform correlations [Geller and Mueller, 1980; Thorbjarnardottir and Pechman, 1987; Harris, 1991]. Interest in earthquakes with correlated waveforms has escalated due to successes in high-resolution relative relocation with correlation picks, for precise delineation of faults and identification of ground truth events for calibration purposes [e.g. Waldhauser and Ellsworth, 2000; Rowe et al., 2002]. The discovery of correlated events for use in relative relocation is now a matter of considerable interest. Schaff, Richards and colleagues [Schaff and Richards, 2004; Schaff et al., 2003] have demonstrated in several regions that a large fraction of all earthquakes occur as correlative twins, driving the potential for widespread application of precision relocation.

Mining explosions also produce highly repetitive waveforms. This fact has been used effectively to classify mining explosions [Israelsson, 1990; Harris, 1991; Riviere-Barbier and Grant,

1993], principally to screen them in test ban monitoring applications and when grooming catalogs of natural seismicity. Commercial explosions also need to be identified to provide training sets in the design and testing of discriminants.

Repeating events pose a challenge to the limited resources of all network operations whether for test-ban monitoring purposes, for hazard assessment or for fundamental investigations of earth structure. Prolific mining explosions and aftershock and swarm events are the principal event types requiring effcient handling [see Figure 1].

Correlation and Subspace Detectors

Correlation-type detectors offer one approach to screening or identifying repetitive events. The classical detector for known signals is the matched filter [e.g. Van Trees, 1968], which correlates a template waveform against a continuous data stream to detect occurrences of that waveform. However, repeating sources frequently produce varied waveforms not well represented by a single template. This situation occurs when events are spread out over a region larger than one or two wavelengths at the dominant frequency of the repeating waveforms [Harris, 1991], or when source mechanisms or time histories are complex and variable. It is desirable to develop detectors that have much of the sensitivity of correlation detectors, but are more broadly applicable.

Current practice in seismic signal detection is concentrated at the extremes of a spectrum of possibilities determined by the amount of information available about the temporal structure of signals to be detected. On one end of the spectrum little signal information is available other than the frequency band where signal energy exists. Incoherent energy detectors, such as STA/LTA algorithms, lie at this end of the spectrum. At the other end, complete information is available about the signal. Correlation detectors, which coherently use the fine structure of signals both temporal and spatial to enhance sensitivity, sit at this end of the spectrum. Each algorithm has advantages and disadvantages. Simple energy detectors are broadly applicable, because they require so little information about the signals to be detected. However, they have high false alarm rates when thresholds are set aggressively to detect smaller signals. Correlation detectors are exquisitely sensitive, having high probabilities of detection at low false alarm rates, even under conditions of threshold detection [Gibbons and Ringdal, 2006]. However, as noted above, they are applicable only to strictly repetitive sources confined to very compact geographic regions.

Subspace detectors span the gap between the strictly incoherent and coherent end-members of this spectrum. These detectors invoke a model that represents the signals to be detected as a linear combination of orthogonal basis waveforms. The essential parameters controlling the tradeoff between sensitivity and flexibility are the detection threshold and the dimension of the subspace, i.e. the number of basis waveforms. In principal, a subspace representation can be constructed supporting a family of detectors that grade nearly continuously between correlation and energy detectors. At the correlation end of the spectrum, the subspace dimension is one; at the energy detector end, the size of the subspace is the dimension of the embedding signal space.

A subspace signal representation is natural when detecting seismic waves traveling a (nearly) fixed path. An individual seismogram $s(t)$ theoretically can be represented by a sum of convolu-

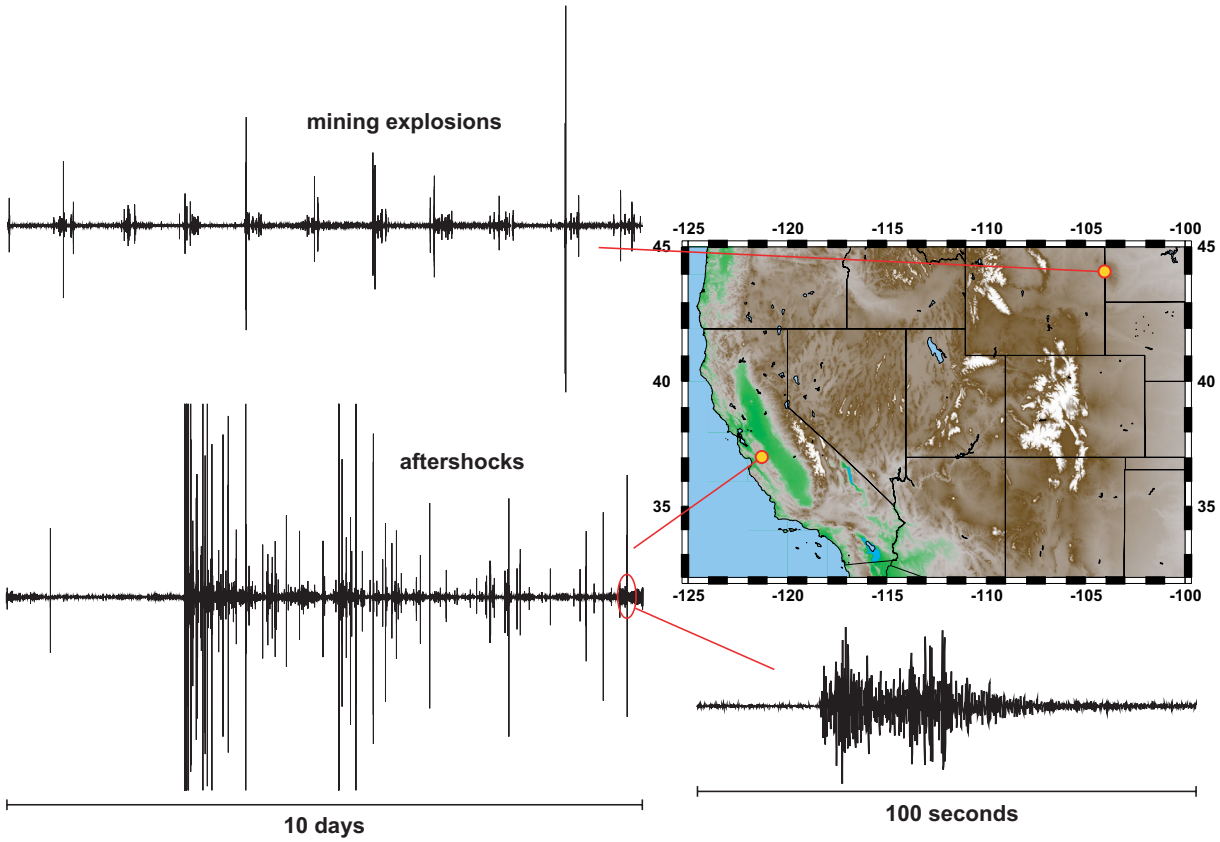

Figure 1 Mining explosions and earthquake aftershock sequences and swarms pose substantial challenges for seismic network operations that may be alleviated with the use of subspace detectors. At top left is a ten-day sequence of waveform data from station RSSD. This station is within 150 kilometers of several tens of large coal mines in the Powder River Basin. The quasi-periodic recurrence of events observed by this station clearly is dominated by the mines. The mines shoot in a regular daily progression of blasts during daylight hours and often at shift changes. Contrasted to this at lower left is the main shock and aftershock sequence of the 2004 Parkfield earthquake observed at station SAO. The events occur at random intervals and are a greater challenge for screening as they are spread over a source region spanning tens of kilometers.

tions of source time histories $h_i(\tau)$ with the corresponding elements of the Green's function tensor describing propagation along the path:

$$s(t) = \sum_i \int h_i(\tau) g_i(t-\tau, \underline{\xi}, \underline{\chi})\, d\tau \qquad (1)$$

For a fixed source location $\underline{\xi}$ and receiver location $\underline{\chi}$ the elements of the Green's function tensor $g_i(t-\tau, \underline{\xi}, \underline{\chi})$ are invariant. Transient events such as earthquakes and explosions have source time histories with finite durations and bandwidths. At any given range, the effective duration of the observed seismogram is finite, for sake of concreteness T seconds. Any signal of duration T seconds with bandwidth B Hertz has at most $2TB$ independent degrees of freedom (Nyquists theorem) and, thus, exists in a vector space of dimension $2TB$. The collection of possible seismograms along a specific path is a subspace of this embedding vector space with potentially far fewer degrees of freedom. Since, in the elastic earth and for sources small enough to be considered point sources, there are 6 independent moment tensor elements, the dimension of the subspace is at most $6 \cdot 2 \cdot T_h \cdot B$ where T_h is the maximum duration of the source time histories. At significant ranges, $T_h \ll T$.

The key to successful exploitation of these observations for efficient detection is to find a low-order basis for the subspace. The approach described in this report is to derive a basis empirically from a collection of representative waveforms characterizing a particular repeating source.

The remainder of this report develops an argument for subspace detectors in five sections. The first formulates the detection problem introducing the new facet of a subspace signal representation, defines a probability model for the observations and derives the structure of the subspace detector. The second section discusses the probability model for the sufficient statistic that the detector calculates, laying the necessary groundwork for optimizing the detector (principally selection of the detection threshold and subspace dimension). The third section presents a procedure for detector subspace design, illustrating the approach using a concrete example of events from the 2002 San Ramon, California swarm. The fourth section measures the performance of the detector against ground truth information available from the NCSN catalog. The paper concludes with a discussion of the implications of the subspace approach (and generalization of the presented approach) for lowering detection thresholds in more general circumstances (e.g. application to arrays) where the signal to be detected is uncertain.

Formulation of the Detection Problem

Our objective is to detect the occurrence of signals of a particular class in a noisy, possibly multichannel data stream. The data may be observations of a single seismic trace, a multichannel waveform from a three-component station or a multichannel waveform from an array or network. We consider discrete-time data streams $x_i[n] = x_i(n\Delta t)$, with n an integer time index. The streams are the digitized representations of bandpass analog signals $x_i(t)$ sampled every Δt seconds. The channel index i assumes values $\{1, 2, ..., N_C\}$ with N_C the total number of data channel streams. For what follows, it frequently will be convenient to pack the N_C individual streams into a single channel-sequential multiplexed stream:

$$x[n \cdot N_C + i - 1] = x_i[n] \tag{2}$$

By convention, when referring to single-channel signals we will use the symbol x_i and when referring to the corresponding multiplexed signals, we will use the symbol x

.

Detectors usually are conceived to implement a binary hypothesis test on the presence or absence of a signal in a data observation window [Van Trees, 1968]. The test chooses between the null hypothesis H_0, that noise only is present, and the alternative hypothesis H_1, that both signal and noise are present (Figure 2). Under several possible criteria, the detection rule is a likelihood ratio test that compares the probability that the observed data are due to signal and noise to the probability that they are due to noise alone:

$$\Lambda(\underline{x}[n]) = \frac{p(\underline{x}[n]|H_1)}{p(\underline{x}[n]|H_0)} < \gamma \tag{3}$$

The data vector $\underline{x}[n]$ reformats into a vector a finite segment of the continuous multiplexed stream of equation 1 (Figure 3) beginning at sample n in the individual scalar signals:

$$\underline{x}[n] = \left[x_1[n]\ x_2[n]\ \ldots\ x_{N_C}[n]\ x_1[n+1]\ x_2[n+1]\ \ldots\ x_{N_C}[n+1]\ \ldots\right]^T \tag{4}$$

The superscript T denotes the transpose operation. Defining the duration of the observation window as N_T samples for each of the individual channels x_i, the total number of samples comprising $\underline{x}[n]$ is $N = N_T \cdot N_C$.

In this discussion, the probability model for the data is multivariate Gaussian. The data consist of noise alone or signal with additive noise:

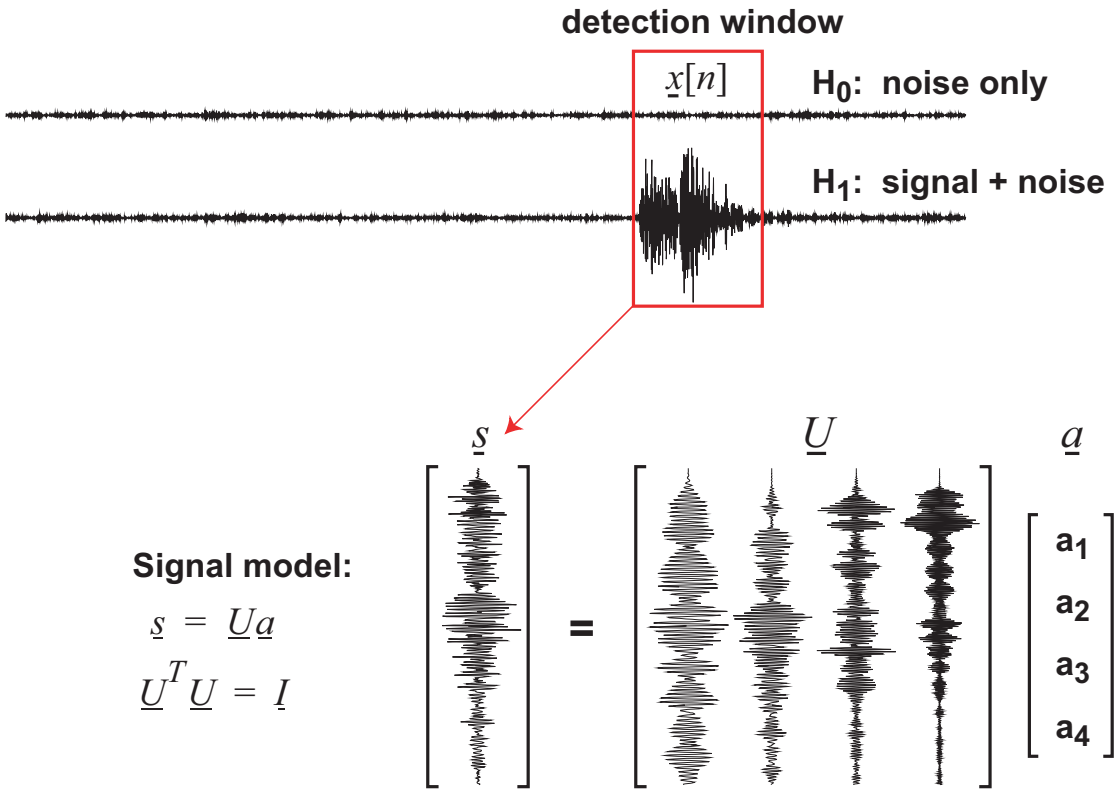

Figure 2 The detection problem is usually formulated for a window sliding along a continuous data stream. At each window position, a binary hypothesis test is conducted, testing whether the data in the window consist of noise alone or signal and noise superimposed. In a subspace detector, the signal is modeled as the sum of weighted basis functions. The weights are considered deterministic, but unknown and must be estimated for each position of the window.

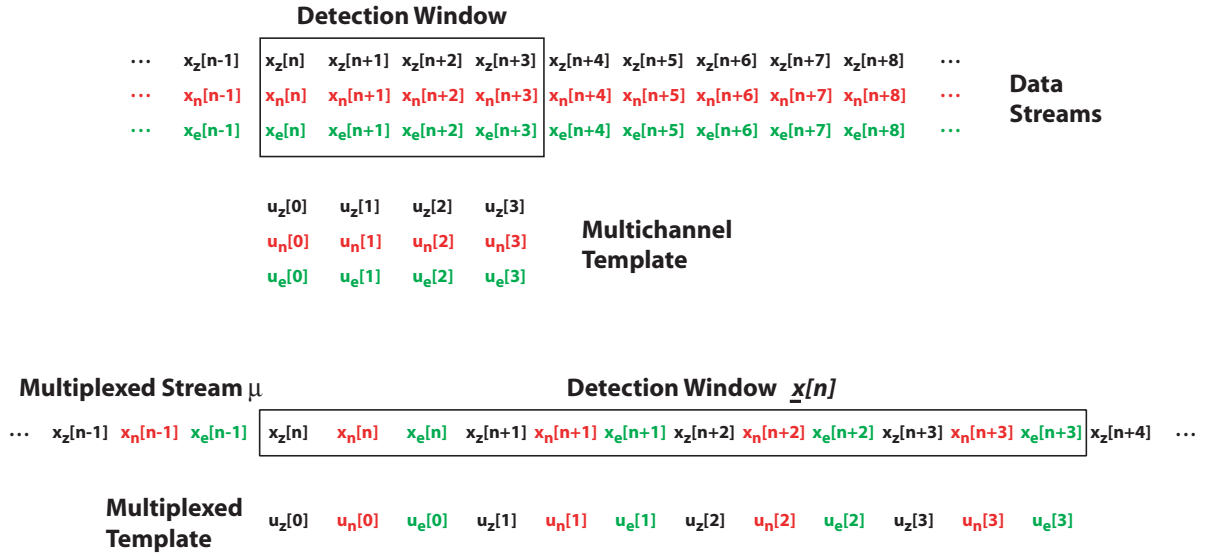

Figure 3 To simplify mathematical expressions and to support an efficient correlation algorithm, it is convenient to reformat multichannel data and signal-matching templates into channel-sequential multiplexed form. The example above shows a portion of a three-component data stream with vertical, north and east (z,n,e) channels and a corresponding three-component template in the upper part of the diagram. In a correlation-matching operation, the template effectively overlays the data in the detection window. The product of data and template samples from corresponding channels (z with z, n with n and e with e) is formed and the resulting products are summed to form the correlation value, as in Figure 1. In the bottom part f the diagram, the same sum of products can be formed by a simple dot product of data and template when reordered into single channel-sequential multiplexed vectors or streams.

$$\begin{aligned} \underline{x}[n] &= \underline{\eta} \quad \text{under hypothesis } H_0 \\ \underline{x}[n] &= \underline{s} + \underline{\eta} \quad \text{under hypothesis } H_1 \end{aligned} \tag{5}$$

The noise $\underline{\eta}$ is assumed to be zero-mean, and temporally and spatially uncorrelated with unknown power level (i.e. unknown variance σ^2). but with the same power on all channels. This simplifying assumption leads to closed-form solution for the detection problem that will be relaxed in follow-on work. The signal is assumed to be deterministic, but dependent upon a vector of unknown parameters $\underline{a}$, and is specified as an unknown linear combination of basis waveforms:

$$\underline{s} = \underline{U}\underline{a} \tag{6}$$

The d columns of the $N \times d$ representation matrix $\underline{U}$ constitute the waveform basis in sampled form. The dimension d may take any value from 1 to N, the number of observed samples. The set of all signals satisfying the form (6) are referred to as the signal subspace. Without loss of generality, $\underline{U}$ can be made orthonormal:

$$\underline{U}^T\underline{U} = \underline{I} \tag{7}$$

where $\underline{I}$ is the $d \times d$ identity matrix.

Under these assumptions, the probability density function for the observed data is

$$p(\underline{x}[n]|H_0) = \left[\frac{1}{2\pi\sigma^2}\right]^{N/2} \exp\left(-\frac{1}{2\sigma^2}\underline{x}^T[n]\,\underline{x}[n]\right) \tag{8}$$

under the null hypothesis (no signal present) and

$$p(\underline{x}[n]|H_1) = \left[\frac{1}{2\pi\sigma^2}\right]^{N/2} \exp\left(-\frac{1}{2\sigma^2}(\underline{x}[n]-\underline{U}\underline{a})^T(\underline{x}[n]-\underline{U}\underline{a})\right) \tag{9}$$

under the alternative hypothesis (signal present).

Because there are unknown parameters in the probability densities under each hypothesis, the likelihood ratio test of equation (3) must be modified to use definite parameter values. A reasonable choice for the parameter values is their maximum likelihood estimates given the available

data. The likelihood ratio test that results is referred to as the Generalized Likelihood Ratio Test (GLRT) [Van Trees, 1968]:

$$\Lambda(\underline{x}[n]) = \frac{\max\limits_{\{\underline{a},\sigma\}} p(\underline{x}[n]|H_1)}{\max\limits_{\{\sigma\}} p(\underline{x}[n]|H_0)} < \alpha \tag{10}$$

The GLRT usually gives good performance in the detectors that implement it.

It is convenient to work with the natural logarithm of the likelihood ratio $l(\underline{x}[n]) = \ln(\Lambda(\underline{x}[n]))$ when the pdfs involved are in the exponential family. Carrying out the maximizations indicated in equation (10) (see Appendix A), the log generalized likelihood ratio is easily shown to have the simple form:

$$l(\underline{x}[n]) = -\frac{N}{2}\left(\frac{\underline{x}^T[n]\,\underline{x}[n] - \underline{x}_p^T[n]\,\underline{x}_p[n]}{\underline{x}^T[n]\,\underline{x}[n]}\right) = -\frac{N}{2}(1 - c[n]) \tag{11}$$

where $\underline{x}_p[n] = \underline{U}\underline{U}^T\underline{x}[n]$, and $c[n] = \dfrac{\underline{x}_p^T[n]\,\underline{x}_p[n]}{\underline{x}^T[n]\,\underline{x}[n]}$

The vector $\underline{x}_p[n]$ is the least-squares estimate of the signal in the detection window, and is obtained by projecting the data in the detection window into the subspace defined by $\underline{U}$ (Figure 4). The quantity $c[n]$ is the ratio of the energy in the projected data to the energy in the original data. It is a positive quantity with values ranging between 0 and 1. It closely resembles the square of a correlation coefficient (it is exactly that for the case $d = 1$, i.e. the correlation detector). This quantity $c[n]$ is the sufficient statistic for a subspace detector.

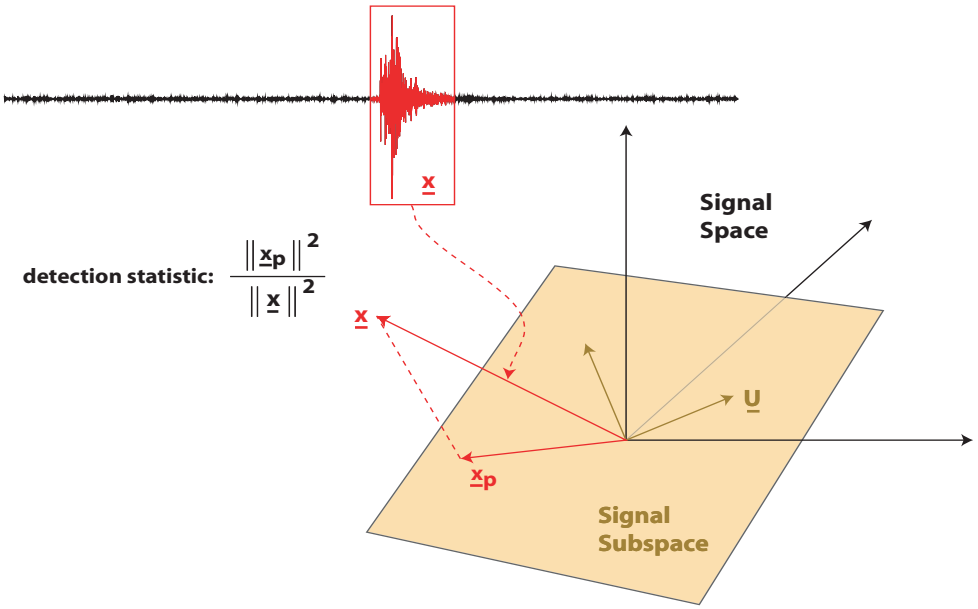

Figure 4 The subspace detector operates by projecting the (possibly multichannel) data in a detection window (treated as a vector $\underline{x}[n]$) into a subspace spanned by the columns of the subspace representation matrix $\underline{U}$. The ratio of the squared norm of the projected vector to the squared norm of the original data vector is the detection statistic. This statistic resembles a correlation coefficient in that it ranges between 0 and 1 and measures the linear dependence between the data and the set of vectors comprising the subspace representation. This statistic is computed continuously as the data window slides down the continuous data stream one sample at a time. When the statistic exceeds a predefined threshold value, a detection is declared.

Theoretical Distribution of the Sufficient Statistic and Detector Performance

Subspace detectors compute the sufficient statistic $c[n]$ (equation 11) continuously and compare it against a threshold, declaring detections when the threshold is exceeded. Estimation and optimization of average detector performance require an understanding of the distribution of the sufficient statistic. Under the simplifying assumption of uncorrelated, white, normally-distributed background noise, the sufficient statistic is a ratio of sums of squares of normal random variables, and, hence, with suitable normalization, has a beta distribution. The numerator of the sufficient statistic, normalized by the noise variance, $\underline{x}_p^T[n]\underline{x}_p[n]/\sigma^2$, is chi-square distributed with d degrees of freedom, since the column dimension of the orthonormal matrix $\underline{U}$ is d. Under H_0 this quantity is a central chi-square variable. Under H_1, when a signal of the form $\underline{U}\underline{a}$ is present, the numerator has a non-central chi-square distribution with non-centrality parameter $\underline{a}^T\underline{a}/\sigma^2$. The non-centrality parameter is the ratio of signal energy to two-sided noise power spectral density [Urkowitz, 1967].

The denominator of the sufficient statistic

$$\underline{x}^T[n]\,\underline{x}[n] \;=\; \underline{x}_p^T[n]\,\underline{x}_p[n] + \underline{w}^T[n]\,\underline{w}[n] \qquad (12)$$

$$\underline{w}[n] \;=\; (\underline{I} - \underline{U}\underline{U}^T)\underline{x}[n]$$

can be decomposed as the sum of two statistically independent terms. The first, the same as the numerator, represents the energy in the data projected into the detection subspace, and the second represents the energy in the orthogonal complement to the subspace. When normalized by the noise variance, both of these terms are distributed as chi-square with d and $N-d$ degrees of freedom respectively.

To compute the probabilities associated with this distribution, it is convenient to manipulate $c[n]$ into a quantity that is F distributed and use available software routines [Mudholkar et al., 1976] for the non-central F distribution. The test for declaration of a detection assumes the form:

$$c[n] \;=\; \frac{\underline{x}_p^T[n]\,\underline{x}_p[n]}{\underline{x}_p^T[n]\,\underline{x}_p[n] + \underline{w}^T[n]\,\underline{w}[n]} < \gamma \quad H_0 \qquad (13)$$

$$c[n] > \gamma \quad H_1$$

As observed above, $c[n]$ resembles the square of a correlation coefficient and, thus, ranges between 0 and 1. Consequently, the threshold γ will be chosen between 0 and 1. Equations (13) can be manipulated into:

$$\frac{(\underline{x}_p^T[n]\underline{x}_p[n]/\sigma^2)/d}{(\underline{w}^T[n]\underline{w}[n]/\sigma^2)/(N-d)} < \frac{\gamma}{1-\gamma}\cdot\frac{N-d}{d} \quad H_0$$
$$\frac{(\underline{x}_p^T[n]\underline{x}_p[n]/\sigma^2)/d}{(\underline{w}^T[n]\underline{w}[n]/\sigma^2)/(N-d)} > \frac{\gamma}{1-\gamma}\cdot\frac{N-d}{d} \quad H_1 \tag{14}$$

Note that the quantity at left in the relations above has a central F distribution with with d and $M + N - d$ degrees of freedom under H_0 and non-central F with non-centrality parameter $\underline{a}^T\underline{a}/\sigma^2$ and the same numbers of degrees of freedom under H_1.

Effects of Signal Mismatch

The discussion so far has relied upon the assumption that the signal, when present, has the form of (6). This assumption implies that the detection subspace has been chosen to represent perfectly the class of signals intended to be detected. As will be made clear in the next section, perfect representation may not be possible or even desirable. The foregoing discussion of the sufficient statistic distribution must be modified to consider what happens when the representation has error, a condition referred to as signal mismatch.

It is convenient to modify the signal representation (6) to incorporate a mismatch term $\underline{Y}\underline{b}$:

$$s = \underline{U}\underline{a} + \underline{Y}\underline{b} \tag{15}$$

with $\underline{U}^T\underline{Y} = \underline{0}$, and $\underline{Y}^T\underline{Y} = \underline{I}$. The signal is acknowledged to have a component in a second subspace orthogonal to the detectors representation subspace. The presence of the mismatch term modifies the distribution of (14) under H_1 to be doubly non-central F (Mudholkar, Chaubey and Lin, 1976) with non-centrality parameters $\underline{a}^T\underline{a}/\sigma^2$ in the numerator and $\underline{b}^T\underline{b}/\sigma^2$ in the denominator. The total signal energy is $E = \underline{a}^T\underline{a} + \underline{b}^T\underline{b}$. The first non-centrality parameter corresponds to the signal energy "captured" by the detector representation and the second to the portion of energy missed by the representation. The fraction of energy captured

$$f_c = \frac{\underline{a}^T\underline{a}}{\underline{a}^T\underline{a} + \underline{b}^T\underline{b}} \tag{16}$$

is a critical design parameter used when optimizing the dimension d of a subspace detector. The dimensions of the representation space ($\underline{U}\underline{a}$) and the error space ($\underline{Y}\underline{b}$) are considered to trade off against each other in discussions of design of the representation later in this report.

Estimate of the Effective Embedding Space Dimension

To this point, the discussion has assumed that the noise in the detection windows is statistically uncorrelated. If the detection window is N samples long, then the dimension of the signal space in which the detection subspace is embedded is also N. As Wiechecki-Vergara et al. (2001) point out, the effective number of degrees of freedom can be significantly fewer if the data are filtered prior to detection. Even without filtering, background seismic noise typically is typically is correlated since seismic spectra are not flat (far from it) and data from three-component and array sensors are spatially correlated. Wiechecki-Vergara et al. provide a means to estimate the effective dimension of the data using the distribution of the sample correlation coefficient $\hat{c}$ computed from a large number of independent, length-*N* samples of background noise. The samples may be obtained by dividing a very long, continuous noise record chosen to be free of transient signals into many hundreds or thousands of length-N data windows.

The Wiechecki-Vergara et al. estimator for the effective dimension of the embedding space in a fixed window is related to the variance of the sample correlation coefficient

$$\hat{c} = \frac{\underline{x}^T[iN]\underline{x}[jN]}{\sqrt{(\underline{x}^T[iN]\underline{x}[iN])(\underline{x}^T[jN]\underline{x}[jN])}} \tag{17}$$

by:

$$\hat{N} = 1 + \frac{1}{\sigma^2_{\hat{c}}} \leq N \tag{18}$$

This value is a conveniently calculated estimate for embedding space dimension for correlated multichannel data such as polarized three-component noise or noise correlated across the spatial aperture of an array. Data samples of the form of equation (4) may be used with many realizations of the sample statistic

$$\hat{c} = \frac{\underline{x}^T[iN]\underline{x}[jN]}{\sqrt{(\underline{x}^T[iN]\underline{x}[iN])(\underline{x}^T[jN]\underline{x}[jN])}} \tag{19}$$

Setting Probabilities of False Alarm and Detection Thresholds

The distribution of the sufficient statistic permits calculations of the probability of detection and probability of false alarm once a detection threshold is chosen. Several philosophies for setting the threshold exist; the one employed here is the Neyman-Pearson criterion (Van Trees, 1968). Under this criterion, the probability of false alarm is fixed at an acceptable value and the probabil-

ity of detection is maximized by adjusting the remaining free parameter: the subspace dimension. The probability of false alarm P_F is determined by the probability that the sufficient statistic exceeds the specified threshold under the null hypothesis. This probability is evaluated from the cumulative central F distribution $F_{d,\hat{N}-d}(X)$ by:

$$1 - F_{d,\hat{N}-d}\left(\frac{\gamma}{1-\gamma} \cdot \frac{\hat{N}-d}{d}\right) = P_F \tag{20}$$

Note the use of the effective embedding space dimension $\hat{N}$ instead of the data window dimension N. The expression of equation (20) must be inverted for the threshold γ for each value of the subspace dimension d. The corresponding probability of detection can be expressed in terms of the cumulative doubly non-central F distribution (Mudholkar, Chaubey and Lin, 1976):

$$1 - F_{d,\hat{N}-d}\left(\left(\frac{\gamma}{1-\gamma} \cdot \frac{\hat{N}-d}{d}\right), \left(f_c \frac{E}{\sigma^2}\right), (1-f_c)\frac{E}{\sigma^2}\right) = P_D \tag{21}$$

where $f_c(E/\sigma^2)$ is the non-centrality parameter for the numerator and $(1-f_c)(E/\sigma^2)$ is the non-centrality parameter for the denominator term.

Empirical Subspace Design and Optimizing Detector Performance

The discussion to this point has presented subspace detectors in abstract terms, describing their form and statistics, but not explaining how a subspace representation can be obtained and optimized. This section presents an empirical approach to subspace design and the use of the statistical model (20 and 21) to optimize performance. This discussion is carried out in the context of an example: detecting events in the 2002 San Ramon, California swarm. The seismological objective of this example is to demonstrate how to obtain a catalog for an earthquake sequence more complete to lower magnitudes starting with relatively few larger-magnitude events.

The 2002 San Ramon Swarm

The 2002 San Ramon swarm occurred in a geographically compact source region (Figure 5), and produced repetitive signals that nonetheless exhibited diversity. Good ground truth information is available for the swarm. The Berkeley Seismological Laboratory (CISN, 2002) analyzed the swarm with the dense Northern California Seismic Network, identifying and locating 179 events in a six day period (Julian days 328-333) and estimating magnitudes. Reported event magnitudes spanned 3 magnitude units (Md 0.71 to ML 3.9), permitting an estimate of detection threshold at certain distance ranges. The largest event was an ML 3.9 strike-slip earthquake that occurred on November 24, 2002 at 14:54:23 GMT.

An excellent selection of high-quality broadband data covering this swarm is available from the Berkeley Digital Seismic Network (BDSN). Two stations were chosen for this study: Kaiser Creek, California (KCC) and Briones, California (BRIB). The two stations are located 238 and 25 kilometers from the swarm, respectively (Figure 5). Three-component data from KCC were used to design, test and run several detectors; data from BRIB were used, along with the Berkeley catalog, to corroborate detections at KCC.

Outline of Procedure

A procedure for constructing and applying an empirical subspace representation, shown in Figure 6, consists of 5 steps.

1. Construct a pool events to use for source characterization. This may be done one of two ways: by extracting events from a catalog, or by processing an interval of continuous data containing the event sequence of interest with an energy (STA/LTA) detector to obtain a pool of larger events for cluster analysis. Typically an STA/LTA algorithm will produce a large number of noise triggers as well as detections of events from the target source and from other sources.
2. Calculate waveform correlations pairwise for all events in the pool, then cluster the events based on the correlation measurements with any suitable clustering algorithm. The waveform correlations for multichannel data can be calculated over all channels of data by forming the required inner products with channel-sequential multiplexed data as in equation (4). The single-link algorithm (e.g. Israelsson, 1990), one of several hierarchical agglomerative clustering techniques, is particularly suitable for the subspace design problem, as will be explained later.

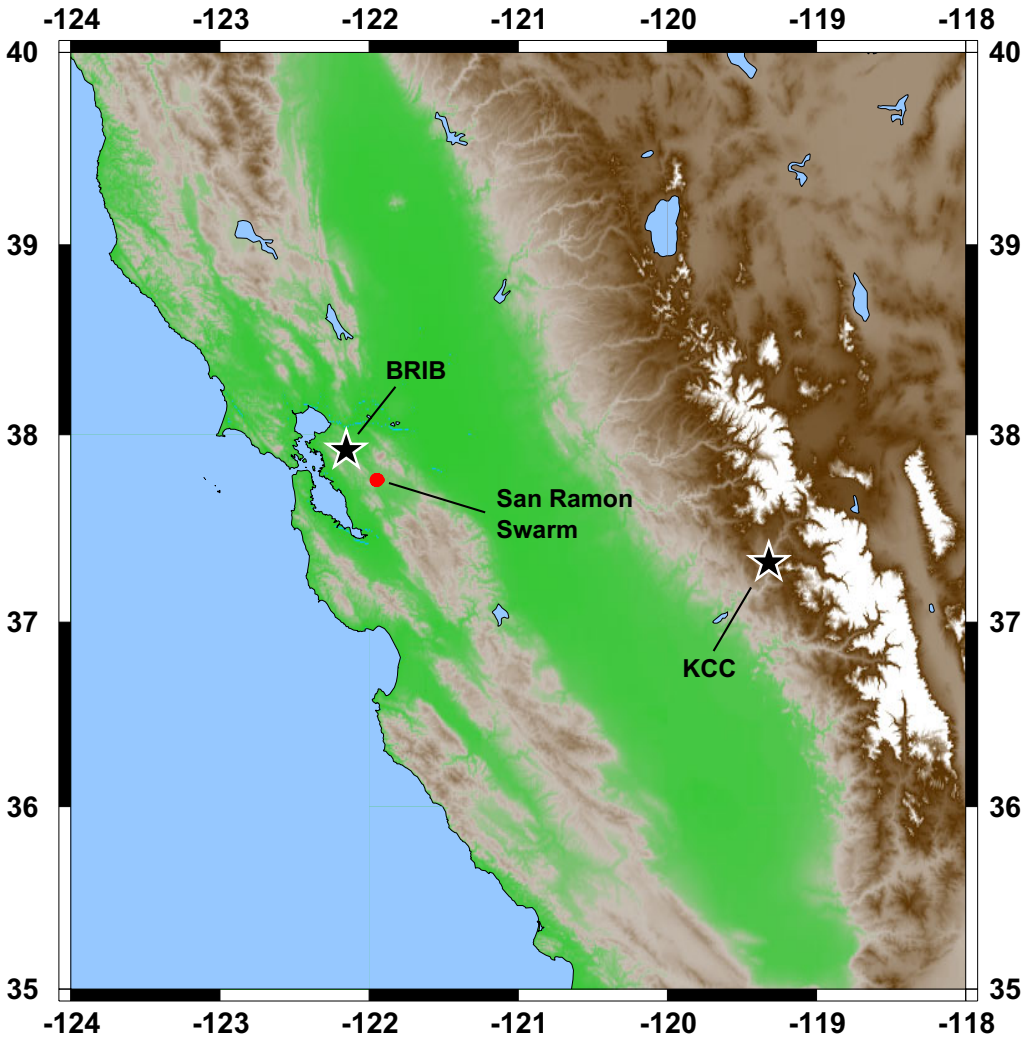

Figure 5 The 2002 San Ramon, California swarm provides a convenient test of subspace detector design and performance. Data from two BDSN broadband stations (KCC, BRIB) are used in the test example, to implement the detector and help assess its performance.

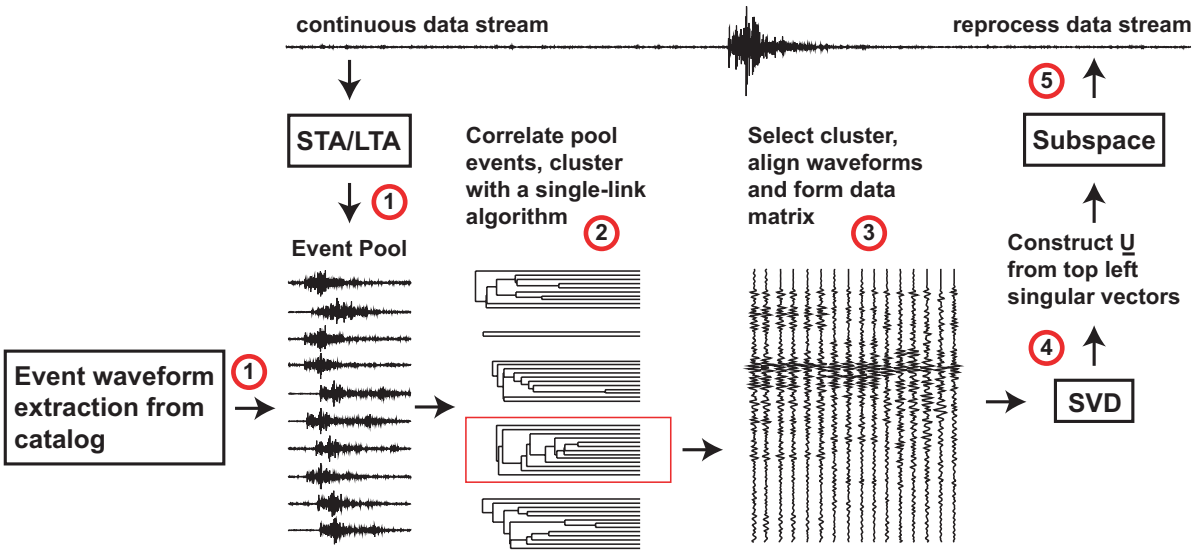

Figure 6 A procedure for empirical design of subspace detectors and for applying them to reduce the detection threshold for a specific source has five steps. See the text for details.

3. Select a (ideally the) cluster associated with the source of interest. The events of this cluster are refered to as the design set. Align the waveforms from the events constituting the cluster and extract the temporal window that will be used to define the subspace template. Proper alignment is crucial; poor alignment will result in a subspace operator with a larger than necessary number of dimensions. If multiple channels of data are available, organize the channels of data for each event as a single vector in channel-multiplexed form (equation 4). Construct a data matrix from the individual event channel-multiplexed data vectors; the event data vectors form the columns of the matrix.
4. Calculate an orthonormal basis for the column space of the data matrix using the singular value decomposition (SVD) algorithm (Golub and Van Loan, 1996). Select the desired false alarm probability P_F and choose the dimension *d* of the subspace detector to maximize the probability of detection over a suitable range of signal-to-noise ratios. The threshold of the detector will be determined by the detector dimension and P_F. As will be described shortly, the singular values hold the key to defining the probability of detection. The subspace representation matrix $\underline{U}_d$ is formed from the left singular vectors corresponding to the largest *d* singular values.
5. Reprocess the data stream with the subspace detector defined by $\underline{U}_d$ to detect events consistent with the design cluster.

Several of these steps require elaboration. The procedure outlined above was applied to three-component data from station KCC spanning the six-day period of the San Ramon swarm. An STA/LTA detector was applied to acquire a pool of 289 events in step 1. A number of issues that arise in executing the remaining steps are illustrated below using the KCC observations.

Choice of Clustering Algorithm

The choice of algorithm for clustering seismic events in step 2 depends on objectives and expectations in characterizing a source. In this report, the assumption is that the source region to be characterized through representative waveforms has some geographic extent, significant variation in source mechanisms, significant variation in source time histories or some combination of all three attributes. An aggressive algorithm for linking events into large chains spanning the space of waveform variations is desirable in this view. The single-link algorithm (see Appendix B) is an appropriate clustering method for this objective. The single-link algorithm aggregates event clusters based only on single pairs of events (one event in each of two clusters under consideration for merging) with a significant waveform correlation. By contrast, a much more conservative alternative, the complete-link algorithm (see e.g. Riviere-Barbier and Grant, 1993), requires all event pairs between two clusters to have significant correlations before the clusters can be merged.

In the case of the San Ramon sequence observed from KCC, a single-link algorithm applied to the 289 three-component waveforms of the event pool produced numerous clusters including one 19-event cluster comprising most of the larger events of the San Ramon sequence. Vertical channel observations for these events are shown in Figure 7.

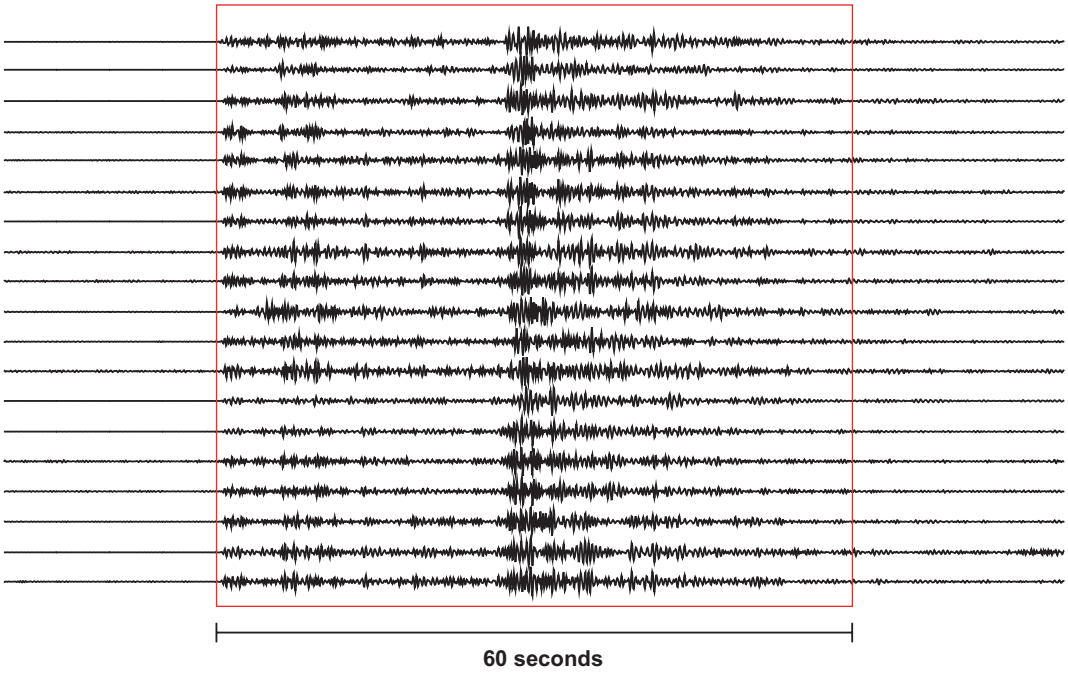

Figure 7 The data window for subspace detector design is chosen to encompass the majority of signal energy. Longer windows only increase processing time with no increase in processing gain. The vertical channel seismograms of the 19 San Ramon events selected for subspace operator design are depicted here. The data window is 60 seconds long and encompasses both P and S phases of the seismograms. Note that, although the signals are broadly similar, they also exhibit significant variation. While only the vertical channel data are displayed here, three-component data were used to form the data matrix in channel-sequential multiplexed form.

Waveform Alignment

Good waveform alignment among the design events in step 3 is critical to obtaining a low-order empirical representation for the signals to be detected. This fact is simple to comprehend by considering the limiting case where all design waveforms are identical. In this case, the dimension of the subspace should be one. However, if waveforms are misaligned, the dimension of the resulting subspace can be as large as the number of unique delays among the collection of signals.

In practice, design events often exhibit substantial variations (Figure 7), which may complicate automatic alignment methods. Typically, these use waveform correlation calculations to establish the optimum alignment delays among the signals [e.g. Rowe et al., 2002]. Waveform variations often result in waveform correlation functions without clear maxima among some event pairs. A solution to the problem of automatic event alignment may be found by using the correlation measurements that do have well-defined maxima. Typically these are the measurements with large correlation values that contribute to link formation in the single-link algorithm used to develop the dendrogram. Consequently, one approach is to align the waveforms through the linkage structure defined by the dendrogram, as described in Appendix B. The waveforms of Figure 7 were aligned using that approach.

Empirical Design of a Subspace Representation

The objective in designing a subspace representation for a subspace detector is to obtain an optimal orthogonal basis for the signals characteristic of events in the target source region. A higher order representation offers the possibility of increasing the probability of detecting a weak event by capturing more of the energy of an incompletely known signal in a correlation operation. However, a representation with larger dimension also will increase the false alarm rate by allowing the detector to match noise with greater probability. Consequently, there is a tradeoff between capturing signal energy and noise energy that must be managed by selecting a suitable subspace dimension. A parsimonious representation consistent with adequate signal energy capture is the objective.

The mechanics of constructing a representation are straightforward. A data matrix

$$\underline{\underline{X}} = \left[\underline{x}_1 \; \underline{x}_2 \; \cdots \; \underline{x}_D\right] \tag{22}$$

is assembled with D multiplexed data vectors (equation 4), one for each event of the design set, constituting the columns of the matrix. In the example followed here, $D = 19$. In order to prevent large events in the design set from dominating the representation, it is convenient to normalize each of the data vectors $\underline{x}_i$ to have unit energy, i.e. $\underline{x}_i^T \underline{x}_i = 1$. The signals are assumed to

have been suitably aligned, perhaps by the algorithm of appendix B. The singular value decomposition [Golub and Van Loan, 1996] of the data matrix

$$\underline{X} = \underline{W}\underline{\Sigma}\underline{V}^T \tag{23}$$

provides an orthonormal basis for the signals of the design set in the form of the matrix $\underline{W}$ of left singular vectors. A smaller basis of any dimension d between 1 and D can be constructed by partitioning the columns of $\underline{W}$ into the group corresponding to the d larger singular values σ_i (diagonal elements of $\underline{\Sigma}$) and a second group containing the remainder of the columns:

$$\underline{X} = \left[\underline{W}_d \middle| \underline{W}_{D-d}\right] \begin{bmatrix} \underline{\Sigma}_d & \underline{0} \\ \underline{0} & \underline{\Sigma}_{D-d} \end{bmatrix} \begin{bmatrix} \underline{V}_d^T \\ \underline{V}_{D-d}^T \end{bmatrix} \tag{24}$$

To the extent that the set of design signals captures the full range of variation in the waveforms from a target region, the signal subspace and the signal mismatch subspace of equation (15) can be identified with the subspace spanned by $\underline{W}_d$ and its orthogonal complement, i.e.

$$\underline{U} = \underline{W}_d \text{ and } \underline{Y} = \underline{W}_{D-d} \tag{25}$$

By neglecting the $D-d$ smaller singular values, the collection of design signals is approximated, in a least-squares sense, as .

$$\underline{X} \sim \underline{W}_d \underline{\Sigma}_d \underline{V}_d^T \tag{26}$$

The matrix of coefficients

$$\underline{A}_d = \underline{\Sigma}_d \underline{V}_d^T \tag{27}$$

can be used to evaluate the energy capture of equation (16), which provides an expression of the fidelity of the d-dimensional representation for the events of the design set. Taking in particular event i:

$$\underline{x}_i \sim \underline{W}_d \underline{a}_i^d \tag{28}$$

where $\underline{a}_i^d$ is the i^{th} column of $\underline{A}_d$. The fractional energy capture (equation 16) for this event is:

$$f_c = (\underline{a}_i^d)^T \underline{a}_i^d \tag{29}$$

the sum of squares of the coefficients. This relation is so simple because the data vectors have been normalized to have unit energy. Another simple relation is the average energy capture for all D of the design events, which can be shown to be:

$$\overline{f_c} = \left(\sum_{i=1}^{d} \sigma_i^2 \right) \Big/ \left(\sum_{i=1}^{D} \sigma_i^2 \right) \tag{30}$$

Figure 8 displays the individual curves (black lines) of fractional energy capture (29) for each of the 19 design events of figure 7 as a function of the dimension d of the representation obtained from the singular value decomposition of the data matrix constructed from these events. The energy capture curves all begin at 0 and end at 1, the latter result a consequence of perfect representation when the dimension of the subspace equals the number of design events. The average energy capture is show as the bold red line in the figure.

Figure 8 is intended as an aid to setting the threshold γ for the detection statistic of equation (13). The detection statistic is itself a measure of energy capture exactly as described in equation (16) when signal only is present in the detection window. Consequently, for the detector to trigger on the design events (a reasonable expectation), the theshold must be set somewhere below the envelope of energy capture curves displayed in the figure. As shown in the figure, a threshold of 0.141 is well below the envelope for a detector with 9 dimensions. In fact, at 9 dimensions, the threshold could be raised to 0.5 without loss of a design event.

Optimizing the probability of detection

A more rigorous approach to estimating detector performance and setting the detection threshold uses equations (20) and (21) with the definition of fractional energy capture established in equation (29). Following the Neyman-Pearson philosophy, the false alarm probability is the starting point, and is set at a value that produces a tolerable rate of false alarms. Once the false alarm probability P_F is set and the number of effective degrees of freedom $(\hat{N})$ of the noise in the detection window is estimated, equation (19) can be inverted for a threshold γ for each of the possible number of dimensions d in the detector subspace representation. With the fraction energy capture (29) then known as a function of d and the threshold γ available, the probability of detection P_D for any particular event in the design set can be evaluated as a function of the signal-to-noise ratio parameter E/σ^2. If it is assumed that (1) the signals in the design set span the range of signals produced by the source of interest, and (2) the design events are all equally likely, then a reasonable approach is to calculate the probability of detection for the source as the average of the individual probabilities of detection for the D design events.

This approach was followed with the San Ramon sequence example with the result shown in Figure 9. In this calculation, the false alarm probability was set conservatively at 10^{-6}. Nineteen individual P_D curves appear in the figure, one for each possible subspace dimension d. The energy capture curves in Figure 8 were used to generate the curves of Figure 9. P_D is plotted as a function of the post-integration SNR (actually $E/(N\sigma^2)$, where N is the number of data samples in the window). The post-integration SNR as defined here corresponds most intuitively to the "eyeball" SNR since it represents the total energy delivered by the signal in the window to the total energy delivered by the noise, and also is equivalent to the ratio of average signal power in the window to average noise power. An SNR of 1 (0 dB) would correspond to rough parity in the average powers of signal and noise.

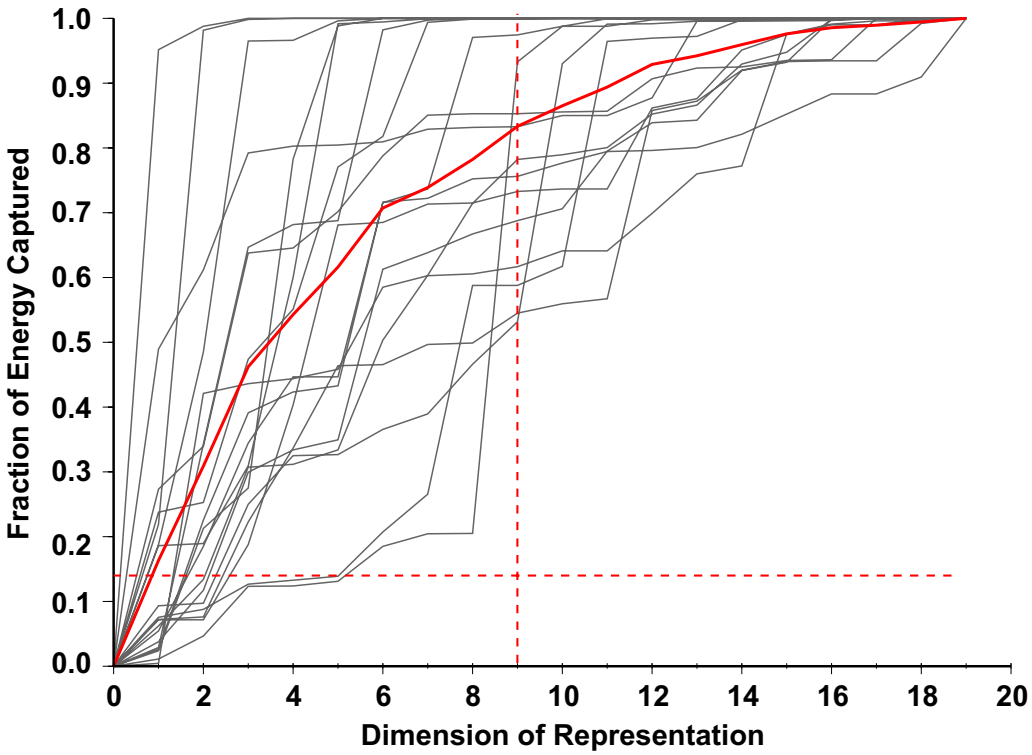

Figure 8 The fractional signal energy capture as a function of representation dimension is an important design parameter driving the calculations of probabilities of detection and false alarm. Each black curve represents the fraction of energy captured in the three-component signal of one of the nineteen design events as a function of subspace dimension. The lower envelope of the curves determines the values of the detection threshold (as a function of subspace dimension) that will result in detection of the design events. The bold red curve is the average energy capture for all 19 events as a function of subspace dimension.

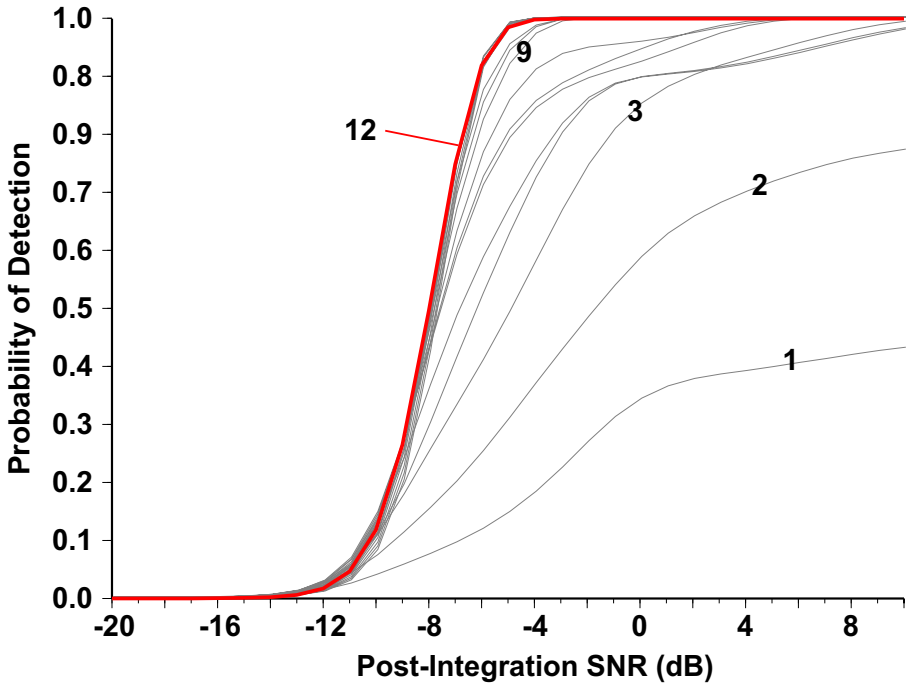

Figure 9 The probability of detection as a function of signal to noise ratio and at a fixed false alarm probability is the most valuable theoretical tool for selecting the dimension of the subspace. In this case, the false alarm probability has been set at 10^{-6}, and the probability of detection calculated as a function of post-integration SNR for subspace representations with dimensions ranging from 1 to 19. In calculating the probabilities of detection, the design events were assumed completely to represent the new events to be detected in the source region (an idealization) and to be equally likely.

Figure 9 shows a pattern common for an event design set with significant waveform diversity. The P_D curve for $d = 1$ does not reach 1 even at SNRs greater than 1 (high SNR). This feature occurs because of the low energy capture that occurs for most events when the representation consists of a single vector (see Figure 8). The energy capture improves quickly and the probability of detection curves climb rapidly as dimensions 2 and 3 are added, but waveforms from two events continue to be poorly represented. The probability of detection curves actually decline somewhat at higher SNRs as 4th and 5th dimensions are added. This feature occurs because energy capture does not improve substantially for the two events, while the ability to match noise increases. Since the false alarm probability is held fixed, P_D must decline when additional dimensions do not improve improve signal energy capture.

By the time the representation grows to 9 dimensions, energy capture is above 0.5 for all events and good detector performance ($P_D \sim 1$) finally is achieved below an SNR of 0 dB. Performance continues to improve, but marginally until 12 dimensions, beyond which the probability of detection curves change little or actually begin to decline. They decline again because the marginal increase in signal energy capture afforded by additional increments to the representation does not offset the increase in noise energy capture. This analysis suggests that a dimension of 9 would be a good choice for detecting the San Ramon sequence events from station KCC.

Observed Performance of Correlation and Subspace Detectors

Figure 10 illustrates the detection statistics obtained by processing the six days of continuous data recorded by station KCC with three different detectors: a correlator constructed from the waveform of the largest (ML 3.9) event in the sequence and two subspace detectors, one with 9 dimensions and one with 12 dimensions. The detectors were designed with template waveforms filtered into the 2-4.5 Hertz band and used all three channels of KCC data (Z, N, E). These detection sequences ($c[n]$ of equation 11) were computed by applying equation 11 repeatedly to data in a window (as in Figure 2) sliding continuously along the data stream shown at the top in Figure 10. The theoretical thresholds computed for a false alarm probability of 10^{-6} are indicated in Figure 10 with the horizontal dashed red lines. The correlation detector has the lowest threshold (0.077) since it has a single dimension and, thus, should match background noise less well than the subspace detectors. It is clear from the figure that the background level of this statistic is substantially lower than the background levels of the other two statistics (which increase with increasing dimension) commensurate with the predicted threshold. It is also clear that the subspace detector have increase processing gain (larger statistics when signals are present) when compared to the correlator.

The thresholds have been set conservatively in this example, resulting in few presumed false alarms, but a number of missed events. The presumed false alarms were those triggers that could not be reconciled against the NCSN catalog events supplemented with detections made by a subspace detector at much closer station BRIB with a conservative threshold. Figure 11 shows the spatial distributions of legitimate San Ramon sequence detections made by the correlator and the 9-dimension subspace detector. At left in the figure are maps (longitude-latitude on top, longitude-depth on the bottom) that show the positions of the events used to design the detectors. The locations plotted in the maps were obtained from the Northern California Seismic Network (NCSN) catalog covering the region and were derived from a dense set of stations operated jointly by the Berkeley Seismological Laboratory and the USGS. The magnitude 3.9 event, used to develop the correlator, is shown as a black star. The other 18 events (black crosses) constitute the remainder of the design set used to construct the subspace detectors. Note that the design events are concentrated in the lower half of the sequence, this being a consequence of the facts that they were obtained by use of an energy detector and the larger events of the sequence are concentrated there.

The events detected by the correlator are shown in the middle of Figure 11. Not surprisingly they are located in the bottom half of the sequence and close to the magnitude 3.9 design event. The events detected by the subspace detector are shown at the right in the figure. They are more numerous and widely distributed as might be anticipated from the more diverse set of design events. Most of the missed events occur in the shallower half of the sequence where the events are smaller and more distant from the design set. In previous studies [e.g. Thorbjarnardottir and Pechman, 1987; Harris, 1991], waveform correlations have been shown to decline substantially with increasing source separation, consistent with the observation of Figure 11.

Figure 12 illustrates the distribution of detected events by magnitude. Three histograms are superimposed in the figure: a histogram (grey) of event numbers by magnitude as reported in the

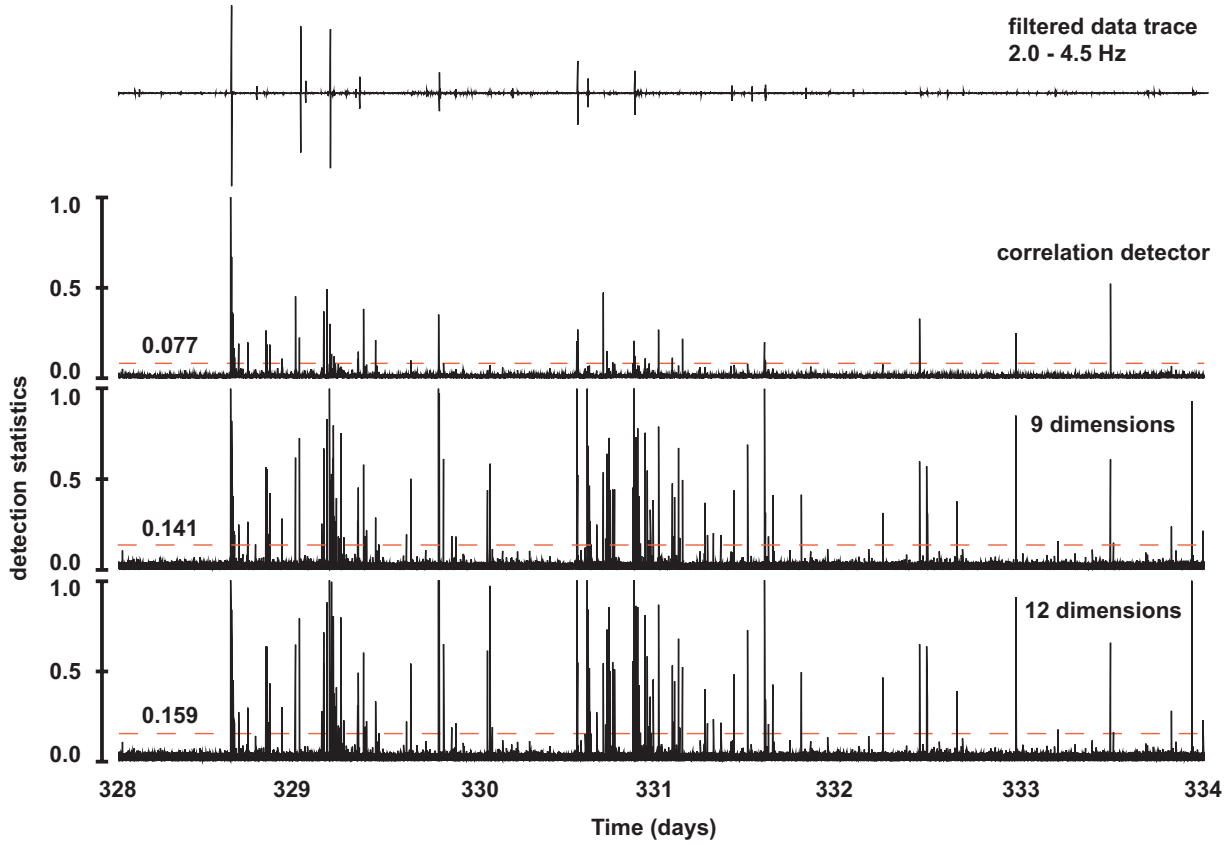

Figure 10 Detection statistics computed for the correlation detector (second trace), 9-dimension subspace detector (third trace) and 12-dimension subspace detector (bottom trace) demonstrate the increased processing gain afforded by subspace detectors. The detection thresholds are shown as red dashed lines and were set in each case to achieve a theoretical false alarm probability of 10^{-6}.

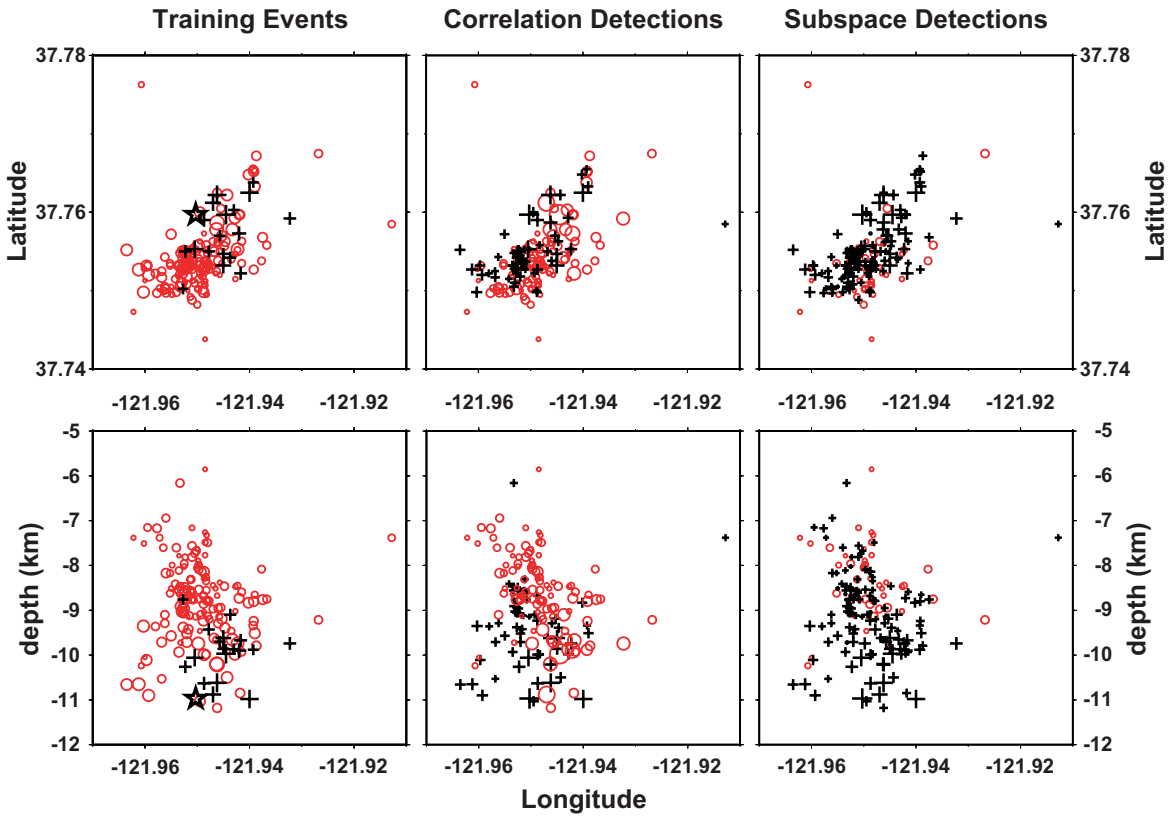

Figure 11 Maps of the locations of events in the San Ramon 2002 sequence as a function of latitude and longitude (top) and longitude and depth (bottom) detected or not detected by the correlation and 9-dimension subspace detector show the greater geographic coverage and sensitivity of the subspace detector. The event locations were obtained from the Northern California Seismic Network (NCSN) catalog and were derived from the dense network of stations operated in the region by the Berkeley Seismological Laboratory and the USGS. The event symbols are scaled to the estimated magnitudes of the events. The main ML 3.9 shock is shown as a star in the maps at left. The crosses in the maps at left are the locations of the design events used to construct subspace detectors. In the center, the maps show the distribution of events detected (black crosses) and not detected (red circles) by the correlation detector using the main shock waveform as a template. At right are the events detected (black crosses) and undetected (red circles) by the 9-dimension subspace detector. The subspace detector detected approximately twice as many events as the correlator using a threshold set to achieve the same theoretical false alarm rate.

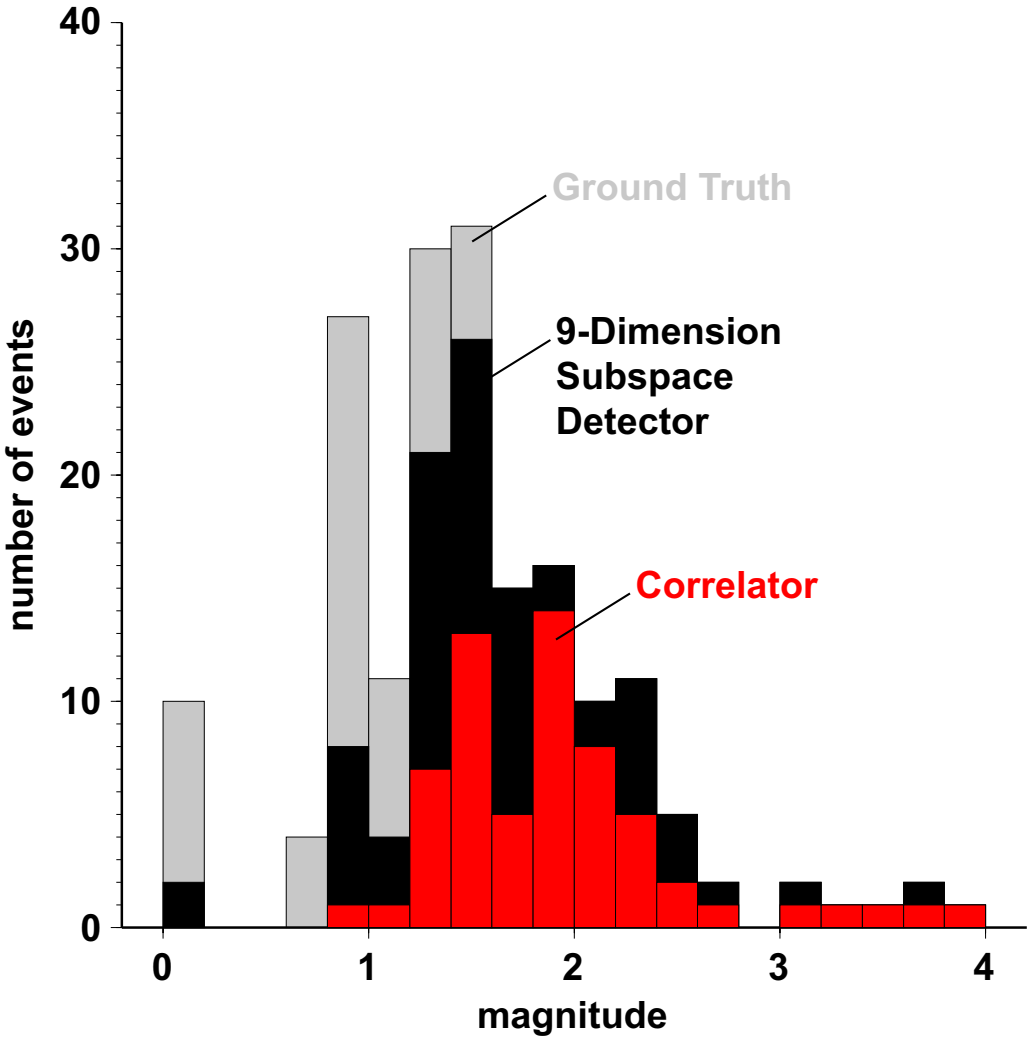

Figure 12 Histograms of the number of events detected by the correlator (red) and the 9-dimension subspace detector are superimposed here on a histogram of the NCSN catalog events by magnitude (grey). The subspace detector captures roughly twice the number of events as the correlator and achieves a detection threshold of about magnitude 1.5 at 240 kilometers range.

NCSN catalog, the histogram for the 9-dimension subspace detector (black) in front of that, and the histogram for the correlator (red) in front of both the other two. The catalog appears to be complete down to about magnitude 1.5, which also is about the threshold of detection for the subspace detector. The correlation detector detects roughly half the number of events as the subspace detector, missing even some of the larger-magnitude events. In this sequence, there apparently is enough signal diversity that the main shock waveform does not provide an adequate representation for all of the larger events.

Subspace detectors also are substantiallly more sensitive for signals from their particular target source than are energy detectors. This point is made clear in Figures 13 and 14. Figure 13 compares the detection statistics for an STA/LTA detector and the 9-dimension subspace detector, both operating in the same frequency band 2-4.5 Hertz. The STA/LTA detector computes the ratio of energies in a 4-second detection window and a preceding 32-second noise window. The log STA/LTA ratio is displayed in Figure 13 and shows a large number of excursions above the threshold (3.0). This threshold was chosen assuming white Gaussian background noise, a central F distribution for the resulting (null hypothesis) statistic and a false alarm probability of 10^{-6}. In fact, there are 735 triggers on the 6 days of data from KCC with 93 being reconciled with the NCSN catalog.

Since the STA/LTA algorithm is a general detector for all types of signals, many of these triggers are legitimate detections of signals from other sources. However, an analysis of the number of triggers and San Ramon sequence detections as a function of threshold suggests that there are a substantial number of noise triggers as well. This analysis is summarized in Figure 14 which shows histograms by magnitude of the San Ramon sequence detections parameterized by detection threshold (values: 3.0, 3.25, 3.5). The total number of triggers are shown in parentheses for each threshold value. The figure shows a rapid increase in the overall number of triggers and in the number of San Ramon sequence detections with only small decreases in the threshold. This behavior is evidence for operation at the threshold of detection, which suggests that large numbers of noise triggers are beginning to occur. The fact that large numbers of triggers are occurring at the threshold predicted for low false alarm probability, is evidence that the assumptions about white Gaussian background noise have been violated.

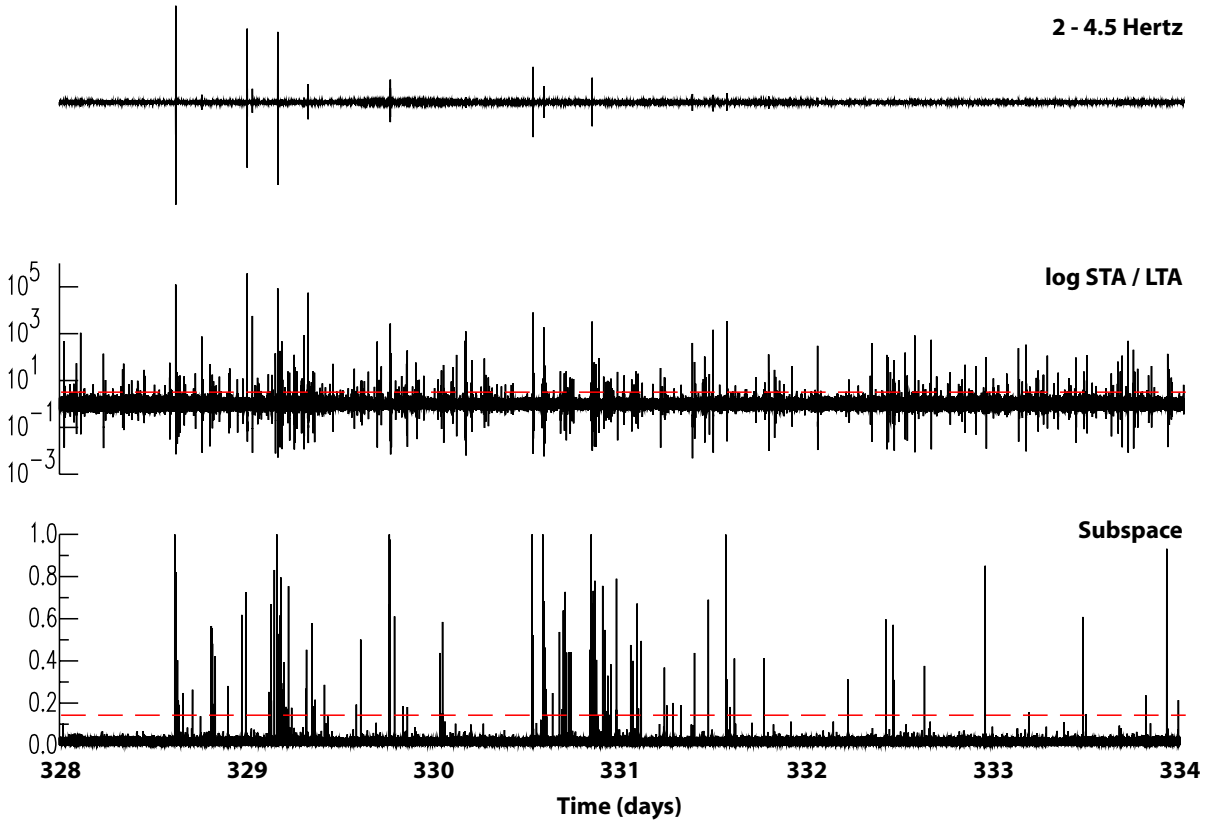

Figure 13 The detection statistic for an STA/LTA detector and the 9-dimension subspace detector demonstrate that the STA/LTA algorithm is not selective for a particular source, as is the subspace detector.

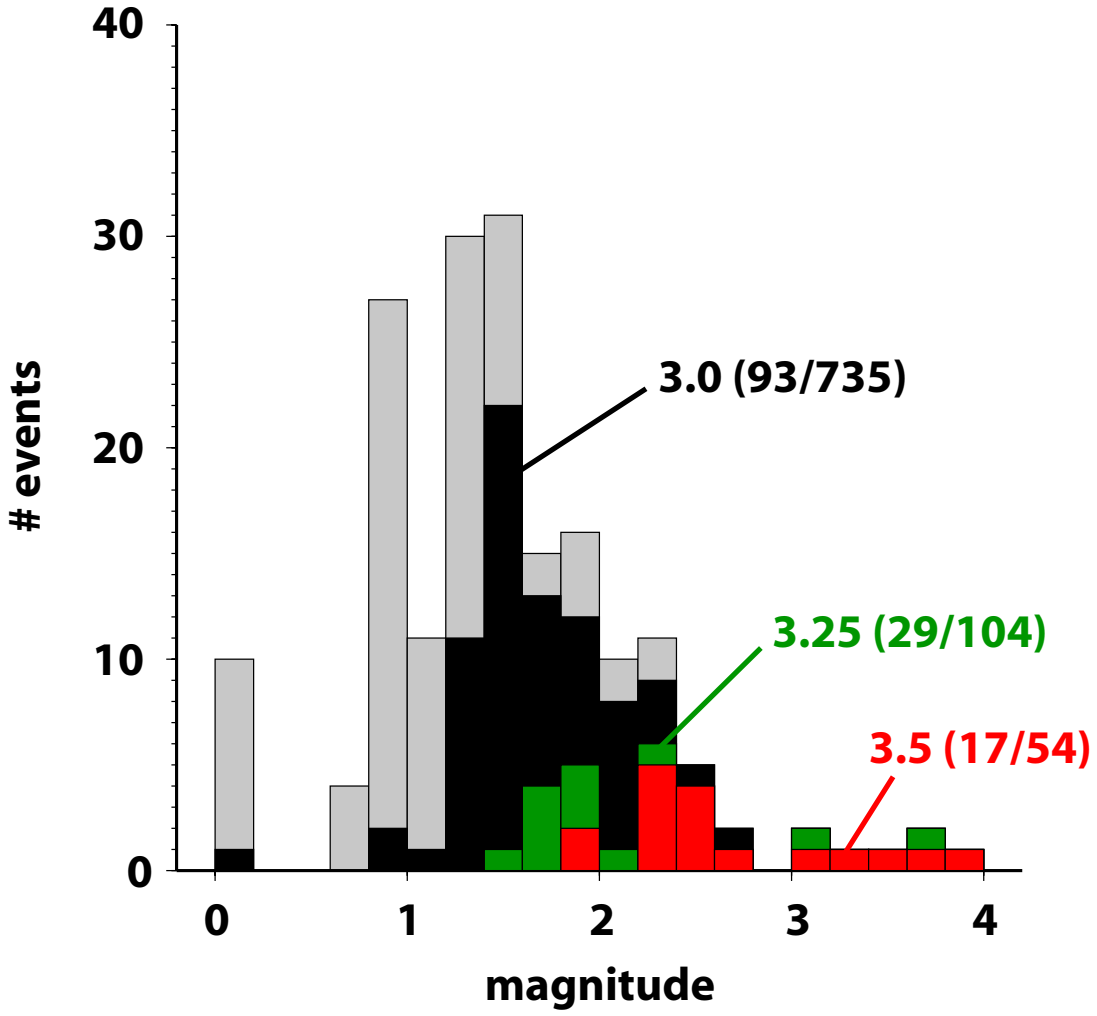

Figure 14 Significant numbers of San Ramon detections with an STA/LTA detector come at the expense of very large numbers of triggers. The figure shows three histograms of the number of San Ramon sequence detections by magnitude as a function of the detection threshold (3 black, 3.25 green, 3.5 red). The number of San Ramon sequence detections and the total number of triggers are shown in parentheses. A detection threshold of 3 is right where the detector exhibits threshold behavior with a rapidly increasing number of triggers (735).

Conclusions

Correlation and subspace detectors operate as sensitive detectors of events in specific source regions because they match the fine temporal and spatial structure of the signals generated by those events, rejecting other waveforms. While the examples shown in this report used multi-channel data from three-component stations, the principle is general and can (and has) been extended to arrays and even to networks of three-component stations. In all these situations, the waveform matching that occurs when a target signal is present and the cross-channel stacking provide processing gain. For a three-component station processing gain occurs from matching the time-history of the signals and their polarization structure.

In an array, this processing gain is similar to beamforming, but with a crucial difference. The fact that the waveforms are matched means that the array sum of single-channel cross-correlations suffers no coherence loss as is the case with a conventional beam of array waveforms dependent on a plane-wave model. The time delays occasioned by wavefield propagation across the array are removed by the channel-wise cross-correlation between identical template and target waveforms, even for arbitrary time delays that do not conform to a plane-wave spatial structure. The fine temporal structure of the signals caused by scattering and refraction similarly is matched. For identical multichannel signals arising from events with the same time history, mechanism and source location, multichannel correlation offers a solution, long sought, to the problem of signal decorrelation across an array aperture [e.g. Mykkeltveit et al., 1983], and leads to highly sensitive detectors [Gibbons and Ringdal, 2006].

However, this attractive solution is degraded if the master event supplying the waveform templates and any event to be detected do not have the same source time history, mechanism or location. The array cross-correlation between otherwise identical events spatially separated declines with increasing distance of their separation [Thorbjarnardottir and Pechman, 1987; Harris, 1991]. This loss of coherence occurs in roughly the same manner, and for the same reason, that the signals observed from a single event decorrelate across an array aperture (in the reciprocal problem). Thus, correlation detectors trade waveform decorrelation problems at the receiver aperture for decorrelation problems at the source end of the path.

Subspace detectors offer one approach to improve, or at least manage, the trade-off. Subspace detectors operate by projecting the data (construed as a vector) from a window sliding over the continuous (multichannel) data stream into the intended signal (vector) subspace. For the detection of seismic events that may be distributed over some source region with a range of source time histories and mechanisms, the subspace can be designed to represent the observed range of waveform variation that reduces correlation detector performance. Thus, subspace detectors provide a mechanism for recapturing some of the performance lost by correlation detectors operating to detect signals exhibiting some variation.

Theoretical estimates of detection and false alarm probabilities can be used to contrast the performance of different processors (single-channel, array, incoherent, and coherent). The suite of probability of detection curves shown in Figure 15 were developed for this purpose [Harris, 2004] from relations similar to equations (20) and (21) for the case where the detection window is supplemented with a pre-event noise window. The figure contrasts curves mapping the probability of

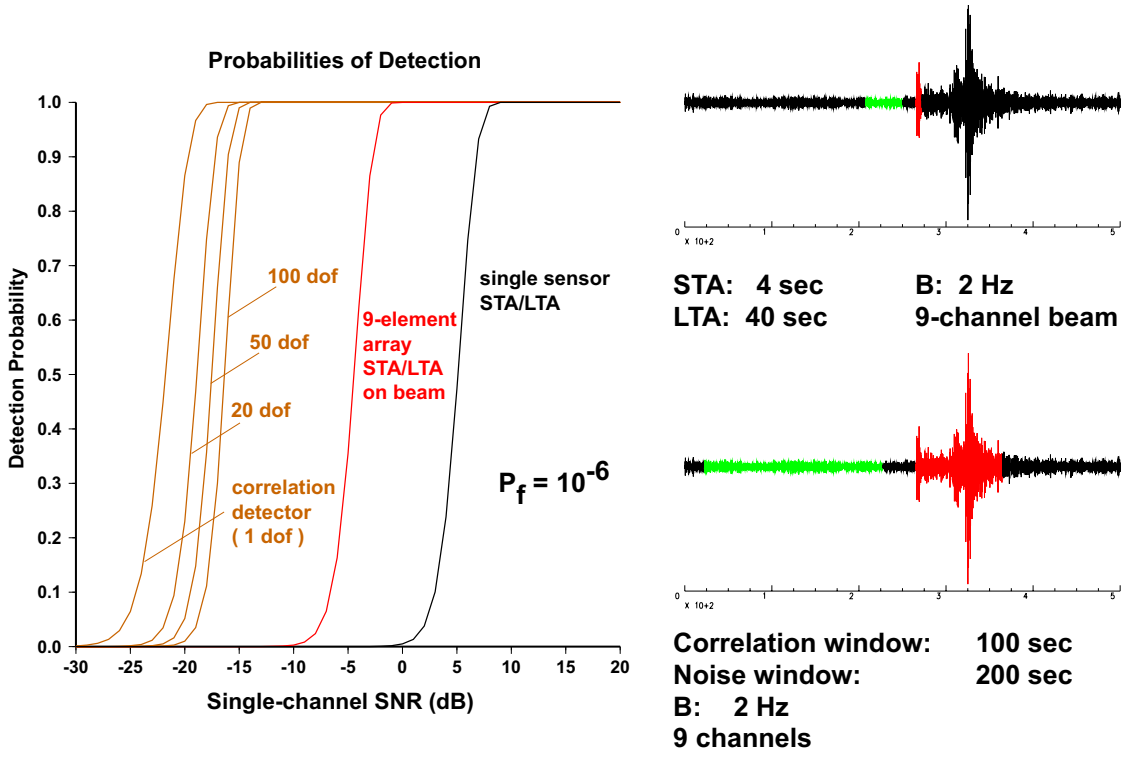

Figure 15 Predicted probabilities of detection for an STA/LTA detector (black), an STA/LTA detector operating on an array beam (red) and four different subspace detectors (gold) suggest that correlation/subspace detectors offer the potential of as much processing gain as the introduction of arrays.

detection P_D for a number of detectors as a function of signal-to-noise ratio (SNR), all for a fixed false alarm probability of 10^{-6} and an assumed processing bandwidth of 2 Hz. One of the detectors is an STA/LTA detector with a short-term detection window of 4 seconds duration and a long-term pre-event noise window of 40 seconds duration. The probability of detection curve for this processor is shown at far right in the figure, predicting a detection threshold in the neighborhood of 5 dB SNR. For reference, a curve portraying the probability of detection for an STA/LTA detector applied to the beam for a 9-element array is shown. It is the next curve to the left and has an apparent detection threshold 9 or 10 dB lower than the STA/LTA detector operating on a single channel of data. The assumptions in making this estimate of P_D are that the noise is incoherent across array elements, white, Gaussian and of the same power on all elements, and the signal is perfectly coherent. From the foregoing discussion, it is apparent that particularly the assumption of signal coherence is flawed. The noise assumptions are highly idealized as well.

To the left of these two STA/LTA curves are four curves for correlation and subspace detectors that also use all 9 channels of array data. The correlation detector P_D curve at far left has a detection threshold that is as much as 17 dB lower than the STA/LTA detector with beamforming. This remarkable increase in performance (decrease in threshold) is partly a function of the increase in the detection window length, and partly due to the coherent integration of energy in a known signal. In correlation or subspace detection, the detection window can be expanded to encompass the entire waveform (all phases) rather than being restricted to a single phase as is common practice in STA/LTA detection on a beam (the restriction is due to beam loss on the secondary phases not the object of beamforming).

A significant assumption in making this estimate of P_D for the array correlation detector is that the signal to be detected is perfectly matched (across all channels) by the template waveform acquired from a previously-observed design event. The noise assumptions are the same as for beamforming. As remarked earlier, the assumption of a perfect match to a prior signal is unrealistic in many cases; the analysis therefore is a best case assessment.

The three remaining curves intermediate between the correlation detector and the STA/LTA detector on a beam are estimates of P_D for three subspace detectors. They are labeled 20 dof, 50 dof and 100 dof for the number of degrees of freedom (dimensions) in their respective subspace representations. The remarkable result of this analysis is that only 5-6 dB of threshold sensitivity is lost with these generalizations on the correlation detector (which is a subspace detector with 1 degree of freedom). The assumptions made in these calculations are that the signal to be detected lies wholly within the subspaces and the noise is uncorrelated et cetera as above. The assumption that the signal lies within the subspace is more likely, of course, than the corresponding assumption for a correlation detector that the correlation template is a perfect match for the signal to be detected.

The theoretical total number of degrees of freedom available in the signal space defined by a detection window of 100 seconds with 2 Hz of bandwidth and 9 channels of data is $2 \cdot 100 \cdot 2 \cdot 9 = 3600$. When contrasted with this figure, even of subspace of 100 dimensions is a small portion of the overall signal space. This fact accounts for the small predicted loss of performance of even the 100-dimension subspace detector. In general, correlation and subspace

detectors perform well because the perform coherent detection of small subspaces of larger signals spaces having $2 \cdot T \cdot B \cdot N_c$ degrees of freedom. This fact motivates the use of large detection windows, as much bandwidth as is observable and as many data channels as are available. Set against this motivation to increase the time-bandwidth-channel product is the fact that signals from events that are separated in the source region decorrelate rapidly with increasing frequency [Thorbjarnardottir and Pechman, 1987; Harris, 1991]. There is thus an incentive to restrict processing to lower frequency bands.

Acknowledgements

The author particularly wishes to thank Dr. Mark Woods for his support and encouragement over the years at critical times in the development of this body of work. The author also wishes to thank Steven Wiechecki-Vergara and Bill Junek for their interest and many helpful conversations and Bill Junek in particular for testing and using the subspace detector software. The author is indebted to Mike Ganzberger, Terri Hauk who supported this work by developing software and providing data used in the development of these algorithms.

This work was performed under the auspices of the U.S. Department of Energy by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48.

References

CISN (California Integrated Seismic Network) (2002), "San Ramon Swarm", http://www.cisn.org/special/evt.02.11.24/

Geller R. J. and C. S. Mueller (1980), "Four similar earthquakes in central California", Geophysical Research Letters, **7**, 821-824.

Gibbons, S. J. and F. Ringdal (2006), "The detection of low-magnitude seismic events using array-based waveform correlation", *Geophysical Journal International*, **165**, 149-166.

Golub, G. H. and C. F. Van Loan (1996), *Matrix Computations*, Third edition, Johns Hopkins University Press, Baltimore.

Harris, D. B. (1991), "A waveform correlation method for identifying quarry explosions," *Bulletin of the Seismological Society of America*, **81**(6), 2395-2418.

Harris, D. B. (1997), "Waveform correlation methods for identifying populations of calibration events," *Proceedings of the 19th Annual Seismic Research Symposium on Monitoring a CTBT*, September 23-25, 1997, 604-614.

Harris, D. B. (2001), "Subspace Techniques for Detecting Repeating Events", poster, SSA 2001 Annual Meeting, San Francisco, CA, 18-20 April 2001, abstract in *Seismological Research Letters*, Vol. 72(2), p 245.

Harris, D. B. (2003), "Detection of Uncertain Signals," *EOS Trans. AGU*, **84**(46), Fall Meet. Suppl., Abstract S22D-01.

Harris, D. B. (2004), "Detection of Uncertain Seismic Signals," 147th Meeting of the Acoustical Society of America, New York, NY.

Israelsson, H. (1990), "Correlation of waveforms from closely-spaced regional events," *Bulletin of the Seismological Society of America*, **80**(6), 2177-2193.

Mudholkar, Chaubey and Lin (1976), “Approximations for the doubly non-central F distributions,” *Communications in Statistics*, **A5**(1), pp. 49-63.

Mykkeltveit, S., K. Astebol, D. Doornbos, and E. Husebye (1983), “Seismic array configuration optimization”, *Bulletin of the Seismological Society of America* **73**, 173-186.

Nadeau, R. M., W. Foxall, and T. V. McEvilly (1995), “Clustering and periodic recurrence of microearthquakes on the San Andreas fault at Parkfield, California,” *Science*, **267**, 503-507.

Nadeau, R. M., and T. V. McEvilly (1997), “Seismological studies at Parkfield V: characteristic microearthquake sequences as fault-zone drilling targets,” *Bulletin of the Seismological Society of America*, **87**(6), 1463-1472.

Riviere-Barbier F. and L. Grant (1993), “Identification and location of closely spaced mining events,” *Bulletin of the Seismological Society of America*, **83**(5), 1527-1546.

Rowe, C. A., R. C. Aster, B. Borchers and C. J. Young (2002), “An automatic, adaptive algorithm for refining phase picks in large seismic data sets,” *Bulletin of the Seismological Society of America*, **92**(5), 1660-1674.

Schaff, D. P. and G. C. Beroza (2004), “Coseismic and postseismic velocity changes measured by repeating earthquakes”, *Journal of Geophysical Research*, **109**(B10302).

Schaff, D. P. and P. G. Richards (2004), “Repeating seismic events in China,” *Science*, **303**, 1176-1178.

Schaff, D. P., F. Waldhauser, P. G. Richards (2003), “Applying massive waveform cross correlation and double-difference location to Northern California and China,” *Eos Trans. AGU*, **84**(46), Fall Meet. Suppl., Abstract S21D-0328.

Scharf, L. L. and B. Friedlander (1994), “Matched subspace detectors,” *IEEE Trans. on Signal Processing*, **42**(8), pp. 2146-2157.

Thorbjarnardottir, B. and J. Pechman (1987), “Constraints on relative earthquake locations from cross-correlation of waveforms,” *Bulletin of the Seismological Society of America*, **77**, 1626-1634.

Urkowitz, H. (1967), “Energy detection of unknown deterministic signals,” *Proceedings of the IEEE*, **55**(4), 523-531.

Van Trees, H. L. (1968), *Detection, Estimation and Modulation Theory*, vol. 1, John Wiley and Sons, New York.

Waldhauser, F. and W. L. Ellsworth (2000), “A double-difference earthquake location algorithm: method and application to the Northern Hayward Fault, California”, *Bulletin of the Seismological Society of America*, **90**(6), 1353-1368.

Wiechecki-Vergara, S., H. L. Gray, W. A. Woodward (2001), Statistical Development in Support of CTBT Monitoring, DTRA report DTRA-TR-00-22, 8725 John J. Kingman Road, Mail Stop 6201, Fort Belvoir, VA 22060-6201.

Appendix A Derivation of the Generalized Likelihood Ratio

Evaluation of the log Generalized Likelihood Ratio

$$l(\underline{x}[n]) = \ln\left(\frac{\underset{\{\underline{a},\sigma\}}{max}\ p(\underline{x}[n]|H_1)}{\underset{\{\sigma\}}{max}\ p(\underline{x}[n]|H_0)}\right) = \underset{\{\underline{a},\sigma\}}{max}\ \ln[p(\underline{x}[n]|H_1)] - \underset{\{\sigma\}}{max}\ \ln[p(\underline{x}[n]|H_0)] \quad \text{(A.1)}$$

is relatively straightforward. The logarithm and maximization operations commute in the above expression because the natural logarithm is monotonic. First consider maximization of the numerator term using equation (9):

$$\ln(p(\underline{x}[n]|H_1)) = -\frac{N}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}[(\underline{x}[n] - \underline{U}\underline{a})^T(\underline{x}[n] - \underline{U}\underline{a})] \quad \text{(A.2)}$$

Setting the partial derivatives of this expression with respect to the unknown parameters (i.e. the partial derivative with respect to σ and the gradient with respect to $\underline{a}$) to zero, we obtain:

$$\begin{aligned} \frac{-N}{\sigma} + \frac{1}{\sigma^3}(\underline{x}[n] - \underline{U}\underline{a})^T(\underline{x}[n] - \underline{U}\underline{a}) &= 0 \\ \underline{U}^T(\underline{x}[n] - \underline{U}\underline{a}) &= \underline{0} \end{aligned} \quad \text{(A.3)}$$

The solution to these equations is:

$$\begin{aligned} \sigma^2 &= \frac{(\underline{x}[n] - \underline{U}\underline{a})^T(\underline{x}[n] - \underline{U}\underline{a})}{N} \\ \underline{a} &= \underline{U}^T\underline{x}[n] \end{aligned} \quad \text{(A.4)}$$

Substituting these values into the expression for the numerator:

$$\begin{aligned} \ln(p(\underline{x}[n]|H_1)) &= -\frac{N}{2}\ln\left[2\pi\left(\frac{\underline{x}^T[n]\,\underline{x}[n] - \underline{x}_p^T[n]\,\underline{x}_p[n]}{N}\right)\right] - \frac{N}{2} \\ \underline{x}_p[n] &= \underline{U}\underline{U}^T\underline{x}[n] \end{aligned} \quad \text{(A.5)}$$

In a similar manner, the denominator term (from equation(8))

$$\ln(p(\underline{x}[n]|H_0)) = -\frac{N}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\underline{x}^T[n]\underline{x}[n] \tag{A.6}$$

is maximized by:

$$\sigma^2 = \frac{\underline{x}^T[n]\underline{x}[n]}{N} \tag{A.7}$$

resulting in:

$$\ln(p(\underline{x}[n]|H_0)) = -\frac{N}{2}\ln\left[2\pi\left(\frac{\underline{x}^T[n]\underline{x}[n]}{N}\right)\right] - \frac{N}{2} \tag{A.8}$$

Assembling the complete result:

$$l(\underline{x}[n]) = -\frac{N}{2}\ln\left[2\pi\left(\frac{\underline{x}^T[n]\,\underline{x}[n] - \underline{x}_p^T[n]\,\underline{x}_p[n]}{N}\right)\right] + \frac{N}{2}\ln\left[2\pi\left(\frac{\underline{x}^T[n]\underline{x}[n]}{N}\right)\right] \tag{A.9}$$

Simplifying:

$$l(\underline{x}[n]) = -\frac{N}{2}\ln\left[\frac{\underline{x}^T[n]\,\underline{x}[n] - \underline{x}_p^T[n]\,\underline{x}_p[n]}{\underline{x}^T[n]\underline{x}[n]}\right] \tag{A.10}$$

which is equation (11).

Appendix B Waveform alignment through the dendrogram

The master event waveforms used to define a subspace detector template should be aligned to minimize the dimension of the basis waveforms constructed to represent them. Often this is a difficult task due to the fact that not all pairs of events may have correlation functions with well-defined peaks suitable for estimating offsets between event waveform pairs. This situation is likely to be encountered, if, as advocated in this report, event clusters are constructed with a single-link clustering algorithm aggressively to form large chains of related events. Waveforms from opposite ends of the chain may not correlate well, posing difficulties for alignment algorithms that impose closure constraints on estimated alignment offsets (e.g. the sum of alignment offsets between events A and B, B and C, and C and A should be zero).

An alternative is not to attempt closure, but rather to rely upon alignments made with the best correlation measurements only (i.e. those with highest correlation values defining offsets). In a single-link algorithm events and clusters of events are aggregated hierarchically using a sorted list of correlation measurements progressing from largest to smallest correlation values. Consequently, this algorithm aggregates events sequentially based upon the largest available correlation measurement at any given point in the execution of the algorithm. The alignment approach used in the examples of this report aligns the waveforms as the dendrogram is constructed using the offsets estimated from the correlation measurements that are themselves used to aggregate events or event clusters. This approach is referred to as alignment through the dendrogram, since the alignments are carried along as the dendrogram is constructed.

A simple algorithm for carrying alignments through the dendrogram is illustrated in Figure C.1. In this example there are six hypothetical seismic events labeled A-F, with correlation measurements in the matrix (upper triangular part) shown at upper left in the figure. In each cell of the matrix the maximum correlation value between the waveforms is indicated along with the delay (in parentheses) between the signals. This delay is the point in the cross-correlation function at which the correlation is maximized. For example, the correlation measurement between events A and B is shown to be 0.9 and this value is measured at an offset between B and A of -50 samples (i.e. the common waveform appears 50 samples earlier in the recording of B than in the recording of A). The event listed along the row of the matrix (A in this case) is considered to be the reference event in the correlation measurement for purposes of establishing time delay.

The single-link hierarchical agglomerative algorithm begins by treating all events as individual clusters contain one event each. In each step of the algorithm, the largest correlation measurement is selected and the two clusters (events) to which it corresponds are merged. As two clusters are combined, the correlation measurements between the two clusters and any third remaining cluster are combined by selecting the larger of the correlation measurements to represent the correlation of the new cluster with the third cluster. This process of aggregation continues until a single cluster remains.

The process is illustrated in Figure C.1. At the outset, in the event correlation matrix at upper left, the largest correlation value is 0.9, between A and B. A and B are merged to form a cluster labeled AB. The correlation between AB and C is the larger of the two values 0.8 (between A and C) and 0.6 (between B and C). The correlations between AB and D, E and F are similarly

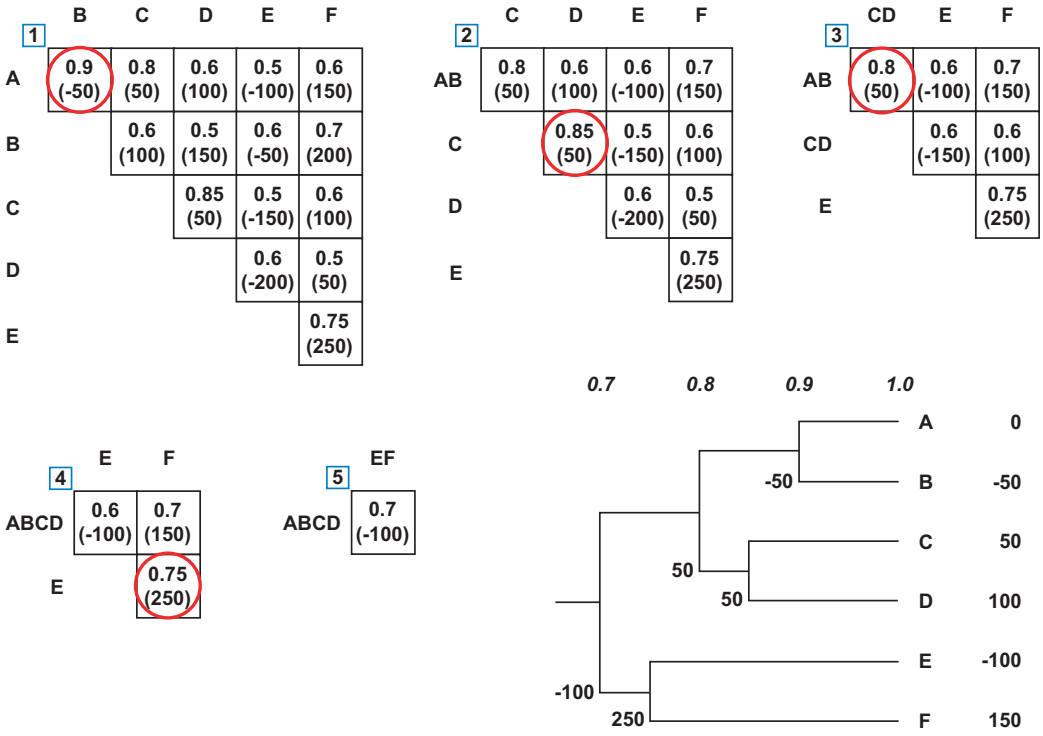

Figure C.1 Waveform alignment through the dendrogram using the single-link algorithm. See the text for a description.

defined. The resulting reduced matrix of event/cluster correlations is shown in the upper middle part of the figure. In the second step, the largest correlation value is 0.85 between events C and D. These are aggregated into a new cluster labeled CD.

At lower right in the figure is the dendrogram representation of the linkage among events and clusters. Events are listed from top to bottom at the right side of the diagram and the links are added backwards from right to left. Events A and B are linked at the 0.9 level. Since B is delayed with respect to A by -50 samples, the branch of the tree leading to B from the node joining A and B is labeled with -50. This value represents the offset required to align A and B. For purposes of timing and alignment each cluster has a baseline event (A in this case) which initially has an offset of 0.

There are several details regarding how offsets are combined as clusters are aggregated. As two clusters are combined, a new baseline event has to be chosen, which is arbitrarily picked as the baseline event of the reference cluster. Subsequently, all measurements with the cluster as reference must have their offsets adjusted to refer to the new baseline event. For example, in the second stage of the example, C and D are aggregated. When C and D are combined, C becomes the baseline event and D has an offset of 50 samples with respect to it. The branch of the dendrogram corresponding to D is labeled with a 50 to indicate the offset with respect to C. Event E has its larger correlation with Event D and an offset of -200 with respect to D. To adjust to the new baseline of cluster CD, E has an offset of $-200 + 50 = -150$ with respect to C, which appears in stage 3. Event F has its higher correlation with Event C and an offset of 100 with respect to C, which remains unchanged.

The process continues until the last correlation measurement is used to combine the last two clusters. At this point all events are referenced to a single baseline event through the linkage structure of the dendrogram. The links are formed sequentially and the accumulating offsets appear in the dendrogram as labels on the branches. To find the aggregate offset representing the delay between a particular event and the baseline event, the offsets are summed down the branches of the dendrogram (starting at the root at left) to the leaf node of the dendrogram corresponding to the particular event in question. In the example, event A ended up being the final baseline event, so that the offsets in the dendrogram (lower right, figure C.1) reproduce the delays of the first row in the initial correlation measurement matrix (upper left).

Figure C.2 shows the dendrogram and the offsets developed to align the 19 master event waveforms recorded at KCC for the San Ramon sequence. Applying the offsets shown in figure C.2 results in the waveform alignments displayed in Figure 7.

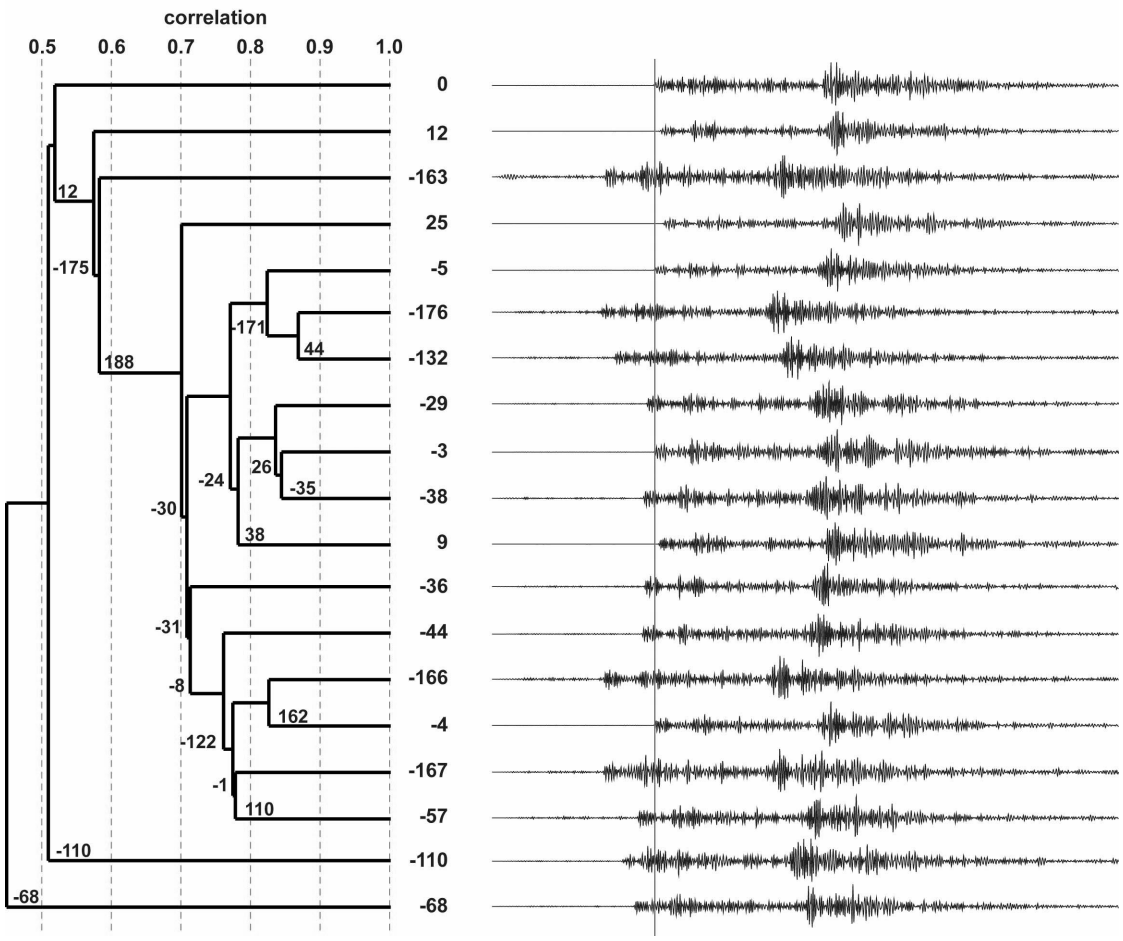

Figure C.2 Alignment through the dendrogram for the 19 master event waveforms recorded at KCC used to build a subspace detector for the San Ramon sequence.